2019

A Comparative Study of Critical Thinking Scores in Teacher Evaluations

Ronald M. Bennett
Long Island University, nesconset@optonline.net

A Comparative Study of Critical Thinking Scores in Teacher Evaluations

by

Ronald M. Bennett

Dissertation Submitted in Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Education

Presented to the Faculty of the

College of Education, Information and Technology

November 2019

Shaireen Rasheed, Ph.D., Dissertation Committee Chairperson

Joseph M. Piro, Ph.D., Dissertation Committee Member

Anthony DeLuca, Ph.D., Dissertation Committee Member

Long Island University

LIU Post Campus

PLACEHOLDER FOR DISSERTATION APPROVAL FORM

ACKNOWLEDGMENTS

ABSTRACT

New York State's Education Law §3012-c (2010) calls for rigorous performance reviews of classroom teachers to assess how curriculum is disseminated in the classroom as part of the educational process. Teacher ratings in New York are derived from a combination of measures, including a state component based on student tests, and a heavily weighted district component that is often more subjective. The current debate about evaluation systems is that student test scores have been used as a measure of teaching abilities that can and has had a detrimental effect on a teacher's career. Because of such a heavy focus on student test scores, parents and several educational groups believe this kind of pressure on teachers is damaging the learning experience for both teachers and students.

This study compared quantitative to qualitative data to gauge discrepancies in scores in the category of critical thinking skills rated categorically by district administrators per the Annual Professional Performance Review (APPR) rubric and how they scored on a self-reporting critical thinking assessment called the Watson-Glaser II Critical Thinking Appraisal. The data verified that categorically rated "Effective" teachers had a higher mean score on the Watson-Glaser Critical Thinking Appraisal than did the categorically rated "Highly Effective" teachers, which suggested a revamping of the kinds of data school districts should be using in the assessment of teacher skills.

Keywords: APPR, highly effective, effective, Watson-Glaser, critical thinking

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER ONE

INTRODUCTION

The U.S. Department of Education has stated there is a growing consensus of concern about America's students and the need for them to be prepared to compete in a world that demands more than the basic skills of reading and writing. Today, about a third of American students require remedial education when they enter college, and current college attainment rates are not keeping pace with our country's projected workforce needs. The National Center for Education Statistics (NCES, 2016) stated that 12 other countries had a 37% higher graduation rate than that of the United States. Countries such as Finland had the highest first-time graduation rate at 63%, Iceland and Slovak Republic at 57%, Poland 50%, New Zealand 48%, Denmark 47%, Ireland 46%, Portugal 45%, Netherland and Norway 41%, Sweden 40%, and finally, Japan 39%. With such a low rating, American educators, governors, business leaders, and parents have called for reforms in education, with specific attention paid to college readiness in an effort to prepare students to compete in a complex world that is globally interactive and steadily increasing in technological advancements.

Background and Context

In order to change the way the United States fares against other countries and their graduation rates, a significant amount of pressure has been placed on high school educators and their efficacy in the classroom. According to studies conducted by the Consortium for Policy Research in Education at the University of Wisconsin, results "positively correlate performance-based teacher evaluation scores with student achievement growth" (Kimball, 2004, p. 54; see also Milanowski, 2004). In New York State, former Governor David Paterson signed Chapter 103 of the Laws of 2010, which added a new Education Law §3012-c with the goal that it

"establishes requirements for new, more rigorous, annual professional performance reviews of classroom teachers and building principals" (New York State Department of Education, 2011, p. 2). The new law now meant teachers and building principals would be under a microscope and need to assess how curriculum and classroom instruction approached meeting learning objectives for students. Further, districts would also have to rethink professional development if the district was falling short in meeting those learning objectives and the national standard. Since the implementation of the new education law, high school teachers now go through a rigorous assessment of their teaching skills to determine how curriculum is disseminated in the classroom as part of the educational process. Aside from domain content knowledge, pedagogical and curricular knowledge, as well as other professional skills, one of the more pressing issues of concern under review is the category of critical thinking and its relation to teacher pedagogy.

Statement of the Problem

In response to creating significant cuts since the recession, 28 states have reduced per-student funding by more than 25% (Hiltonsmith & Draut, 2014). A competitive edge was created by many states when the federal government began to roll out incentives for increased student achievement levels in all academic areas as they directly relate to the ever-increasing need to produce students who are college-ready. However, the pressure of increasing student achievement lies primarily on the shoulders of teachers, as they are the ones responsible for the transmission of curriculum. Teaching quality is a key factor influencing student outcomes (Odden, Borman, & Fermanich, 2004). In an effort to continue meeting local and national standards of student achievement and assessment, the evaluation of teacher skills has become a critical component of not only student success measurement, but also a school district's ability to hone in on how teachers are implementing the curriculum based on individual teacher skill sets.

The most challenging aspect of evaluating teachers is if new systems or models of evaluation incorporated in school districts statewide can provide accurate results in efforts to pinpoint where a particular teacher may be lacking skills in his or her traditional pedagogy. Rothstein (2010) would agree, suggesting that because demographics may vary from one class to another, evaluations should be done quantitatively to account for student body makeup in the district's area. Additionally, teachers should be assigned to classes with a clear understanding of how the demographics of the classroom impact student learning, especially in cases of special education, English language learners, and students in gifted programs. It may be tempting to assign students within these categories to specific teachers, but not all districts may have that capability. School districts are more concerned about acquiring monies to improve their districts, which means aligning themselves with standards reported by "successful" school districts with test scores that could set a national average.

The 2012 *Race to the Top* (RTT) federal grant process required states to redesign evaluation systems that endorsed effective teaching and integrated student achievement data in educator evaluation systems. RTT was designed to have exponential results for teachers, students, and districts. The grant's purpose, offering a payout of $4.35 billion, was to "reward states for past accomplishments, create incentives for future improvements, and challenge states to create comprehensive strategies for addressing the four central areas of reform that will drive school improvement" (U.S. Department of Education, 2009, para. 1). Those areas of reform are: (a) designing and implementing rigorous standards and high-quality assessments, (b) attracting and retaining quality teachers and leaders, (c) supporting data systems that inform decisions and improve instruction, and (d) using innovative reforms to transform struggling schools (U.S. Department of Education, 2009). The grant further purported to encourage districts to adopt

standards that would prepare students for success in college and the workplace; improve at-risk schools; improve teacher recruitment, professional development, and retention; and reward "effective" teachers and principals. These parameters called for a major reform of educational practices that would continue to impact education over the next decade.

However, the one item of reform that has consistently been under public scrutiny is the call to "measure student success and inform teachers and principals about how they can improve instruction" (U.S. Department of Education, 2009, para. 1). More importantly, there were concerns regarding the efficacy of new teacher evaluation designs and their ability to measure a teacher's ability accurately based on students' test scores. An additional criticism was the cost to districts to apply for the grant. An article by Annie Hsiao (2011) written for the *National Review* stated, "In addition to RTT's few and limited results, GAO reports that applying to RTT took thousands of hours and additional staff. State officials said they spent $75,000 to $620,000 on hiring application consultants. It may simply be too soon to tell just how effective, if at all, RTT will be" (para. 6). Hsiao's assessment hits on several sensitive points that are challenging to all districts—funding and implementation. A larger concern that has been expressed by parents and teachers alike is that the grant would cause districts to prioritize test scores over the teacher and student learning experience.

The proposal of any new system, no matter how ultimately successful, faces complicated opposition that can end up doing more harm than good. We know this from Presidents Bill Clinton and George W. Bush with the No Child Left Behind (NCLB) legislation, which was the first reform since President Lyndon Johnson's Elementary and Secondary Education Act (ESEA) of 1965. According to Fritzberg (2012), "Presidents Bill Clinton and George W. Bush attempted to bridge the concerns about both quality and equality in public education through promoting

statewide standards and assessments that all children should achieve" (para. 5). However, not all districts in one state have a unified demographic, which is why President Barack Obama's form of educational reform also wanted to focus on how districts perform by looking at the demographics of race, ethnicity, and socioeconomics. Coincidentally, one of the results of Clinton's and Bush's NCLB initiative was instituting teaching standards in the category of being "highly qualified." "Highly qualified" teachers had to hold a bachelor's degree and a state license, as well as demonstrate competency in their subject matter (U.S. Department of Education, 2009b). Better qualified teachers will produce positive student outcomes, but there is now an overemphasis on highly qualified teachers based on test scores.

In May 2010, the New York State Legislature tried to ensure that the RTT program adopted an amendment to Educational Law 3012-c regarding the Annual Professional Performance Review (APPR) of teachers and principals. The new amendment meant teachers would now have a numbers-driven incentive to produce test results that met state standards. Most, however, did not perceive this amendment as an incentive, but rather as the first step in negatively affecting classroom instruction and ethics. In an open letter opposing the new APPR ruling prepared by the President of the Nassau County High School Principals' Association, Sean Feeney (2013) explained why the new ruling is problematic: "The new law states that beginning September 2011, all teachers and principals will receive a number from 0-100 to rate their performance. Part of that number (ranging from 20% to 40%) will be derived from how well students perform on standardized tests" (para 4). Feeney went on to list three major concerns regarding the impact on students and teachers. According to the letter, Feeney asserted that this new law will negatively impact students because it will cause a shift in teacher priority—especially if a teacher must shift focus onto student scores on standardized tests because it will

directly impact their livelihood and career as well as student-centered engagement. While test preparation is important, the shift in priority will take away focus from other important factors of student learning, such as student enrichment programs.

Unfortunately, Sweeney's concerns would later come to fruition. In a survey published by *Newsday*, "nearly 65,000 students in Long Island elementary and middle schools refused to take English Language Arts test…100 of the island's 124 public school districts, 64, 785 of 148, 564 children opted out of the exam" (Tyrell, 2016, para 2). The numbers may have been worse than that, as some districts did not want to divulge exactly how many students opted out. There is a real possibility of the number being double of what was reported in this one survey.

While the trend of educators and parents coming together in protest of the tests is growing, some feel that opting out hurts not only the districts and teachers, but also the students. As Jonathan Burman (2015) of the New York State Education Department (NYSED) said, "Test refusal is a mistake because it eliminates important information about how our kids are doing. Those who call for opting out really want New York to opt out of information that can help parents and teachers understand how well their students are doing" (para 12). Nicole Brisbane (2015), state director at Democrats for Education, agreed with Burman: "Collecting educational data is important for the future of education and can help define the character of a town" (para 19). Ironically, data collection is at the crux of tension over testing and teacher evaluation. If a large number of students are opting out with parental support, the districts must look at what other factors are involved in that decision.

It is important to note that numbers reported for opt out were specific to Long Island school districts and considered a movement belonging to upper middle-class suburbia. Statistics shared by NYC Opt Out (2017) presented data reflecting large numbers of students who opted

out of testing belonging to students who were labeled as "economically disadvantaged." According to NYC Opt Out numbers derived from NYSED's District-Level Test Refusal File, school districts in upstate New York and NYC make up at least "45% of New York State's public school students" (para 2.). The report further explained that in NYC, "60% of children who opted out of ELA were economically disadvantaged, 47% of children were students with disabilities, and over 12% are English Language Learners" (NYC Opt Out, 2017) (see Figure 1.1).

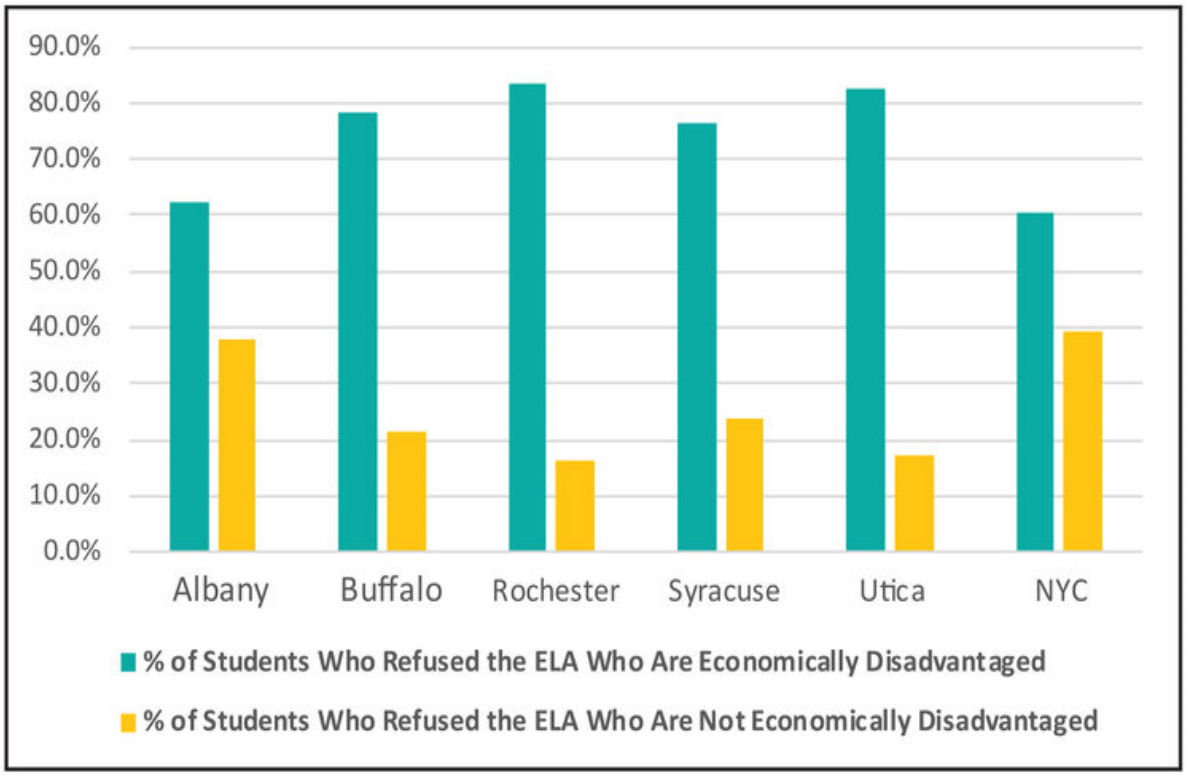


Figure 1.1. Opt-out trends of economically disadvantaged students across New York State (NYC Opt, 2017)

That 60% of children who opted out accounts for a large number of students who face challenges that suburban kids might never have to face. Therefore, how does teaching instruction measure up to children dealing with language barriers, cognitive delays, and other environmental issues that encumber the learning process? Further, how can a teacher achieve accurate ratings in his or her success as an instructor when students wrestling with these challenges struggle to overcome their learning difficulties with the best of teachers and still do poorly on the exam? A number of factors impact the learning process for economically disadvantaged students like poverty, abuse, neglect, drug abuse, and addiction, the fracturing of the family unit, and cognitive delays that may go untreated because of a lack of access to resources. One of the issues with how the current teacher evaluation system works is that the assessment is not calibrated to consider the makeup of the actual class being observed. Regardless of the makeup of the class, the teacher is still expected to produce scores from students that demonstrate successful teacher instruction.

The pressure on teachers from the district is also complicated. If the data regarding the number of students who participated in the standardized tests fall under a 95% participation rate, that could mean a significant amount of funding will not go to the state, which filters down to the districts. The need to maintain this participation rate is an additional layer of responsibility that not only affects teachers but also school administrators who are tasked with trying to reach and maintain a certain standard. This dynamic has proven to be problematic for several reasons, the least of which is the undue pressure teachers feel to perform at a truly unquantifiable capacity. Dependence on test scores also does not account for differences in socioeconomics when it comes to poorer neighborhoods where schools are severely under-financed for programs that address the remedial needs of students. A study by Daniels (2013) titled "APPR, Solution or

Problem" stated, "third-party suppliers are offered the potential to create evaluative rubrics (which require training and potentially retraining) that may generate substantial amounts of money for those companies or organizations that submit successful applications and thus become a part of the approved list" (p. 25). There may, in fact, be positive attributes for revitalizing and reassessing standards in the present education system, especially in the area of student achievements necessary for college preparation. However, there is also a need for an examination of the administrative observation criteria during the assessment of a teacher's skills in the annual performance review process. This examination is critical in order to measure accurately if an employee's teaching skill matches up with the implementation of such a large educational shift such as the relatively new and still controversial Common Core Standards.

Professional Development

Professional development must also be examined in order to support the evidence of a direct correlation between student learning and teaching skills, as specified by the criteria of Pearson's Annual Performance Peer Review. Since the inception of the APPR, teachers have been attempting to understand the positive effects of the evaluation system on student achievement. The concern, however, is "while more effective hiring and firing practices may increase average teacher effectiveness over time, it fails to address the majority of teachers who are currently in classrooms" (Maharaj, 2014, p. 2). Therefore, adjustments need to be made in the assessment of teachers who have been in the education system for several years and who may have had different training than newer teachers of the current time. This kind of alignment would go far in norming expectations across the gap between newer and more seasoned teachers.

By investigating quantitative discrepancies of APPR evaluation scores, such as the critical thinking component, valuable information garnered from that kind of assessment would

be useful in identifying the misalignment between the design of the evaluation system and comprehending the tasks of the teaching practice. Understanding any discrepancy in how teachers are evaluated is also valuable in how districts approach professional development, which directly feeds into how teachers continue to develop their personal pedagogy in an ever-changing educational system. Smylie (2014) stated, "One factor most consistently associated with the lack of impact is the troublesome relationship between evaluation and professional development—the opportunities for teachers to learn and to improve their practice in response to and beyond the process of evaluation itself" (p. 98). Therefore, strong, high-quality professional development is necessary in order for teachers to improve their practice.

A national study of 1,000 teachers across the nation reported that only one quarter of teachers considered their recent formal evaluations valuable and effective (Duffet, Farkas, Rotherham, & Silva, 2008). With the onset of the new evaluation system having had an obvious impact on the perspective of teachers in New York State, it is important that the educational system and teacher evaluators provide a reliable and explicit willingness to align data in the areas vital with student achievement and college readiness after high school. There are well-known concerns that followed the initial implementation of APPR. Other general disputes about APPR are the Value-added model (VAM) used in the evaluation of teacher effectiveness, which does not necessarily create stable ratings of teachers. In essence, different statistical models yield different effectiveness scores (Papay, 2011), which also problematizes the accuracy of teacher effectiveness assessment, as a teacher's rating changes from class to class, from year to year, and even from test to test (McCaffery, Lockwood, Koretz, & Hamilton, 2003). Because of these fluctuations, I argue evaluation systems need to be tested for consistency and reliability, especially in the realm of assessing a teacher's effectiveness. As Vansickle (2012) stated:

> Dewey claimed that a person must have certain attitudes in order for reflective thought to occur. Measures are needed to assess open-mindedness, intellectual self-confidence, willingness to postpone judgment, willingness to test one's beliefs, valuation of knowledge and thinking, demands for closure, and desire for intellectual consistency. (p. 9)

Most of these attributes are aligned with a teacher's need to fundamentally review and evaluate his or her own criteria for critical thinking skills, along with being judged fairly and accurately by school district evaluators. "Self-evaluation is a potent learning incentive and a procedure too seldom exploited" (Dressel & Mayhew, 1954, p. 20). However, in order to measure teacher effectiveness accurately and allow teachers more agency in determining how to impact student success, the evaluation system should be structured to include input from teachers who are on the frontline of curriculum implementation. Granting teachers this kind of input gives them a central role in decision making within the instructional and managerial framework of the school system and positively impacts professional development.

A study by Sagnak (2010) examining the relationship between transformational school leadership and ethical climate found that there are positive outcomes when an organization invokes participative leadership that influences shared responsibility between superiors and subordinates. Sagnak also mentioned that participation in decision making contributes to the quality of work life and improves professional training. This would mean school district administrators would use teacher input as valuable data in terms of professional development needs, assessment models, and curriculum like that of the Common Core Standards. Assessing teachers' perceptions toward evaluation methods by surveying explicit questions regarding its effectiveness is useful information in exploring and aligning teachers' opinions toward student

education. According to Dr. Nathanial Schwartz (2013), director of the Office of Research and Policy at the Tennessee Department of Education, "A final important lesson is that teachers who perceive the system as focused on teaching improvement rather than judgment about their performance tend to engage with and value teacher evaluation to a far greater extent" (para. 11). In that vein, the Tennessee Consortium on Research, Evaluation, and Development has had an active approach to satisfying teachers' perception levels to their evaluation system. The Tennessee Department of Education has offered a survey to test teacher perceptions pertaining to evaluation systems, with specific questions that help to align ideas indicative of student achievement levels. Tennessee's DOE analysis of teacher perceptions was completed by 25,000 teachers and 3,000 administrators, representing 39% of teachers and 46% of administrators, respectively, across the state (Schwartz, 2013, para 12).

The Impact of Teacher Evaluations and Subsequent Response

A rating scale incorporating four possible characterizations describing the performance level of teachers, as former New York City School Chancellor Carmen Fariña (as cited in Decker, 2014) explained, is "A well-developed evaluation system—with four, much more nuanced ratings, instead of only two—helps us identify and provide specific support to struggling teachers, as well as identify those who do not belong in the classroom" (para 10). Such an active approach to make teacher evaluations more useful at all performance levels should be used by many districts and states, especially where the teachers may have the opinion that the evaluation system, of which 20% is based on student scores, was rushed and had a disastrous implementation process. This method of scoring is coming into question by lawmakers; however, the Preliminary Statewide Composite HEDI results indicated that 94% of teachers and 92% of principals obtained ratings of both "Highly Effective" and "Effective."

Decker (2014) stated that "Cuomo has railed against the current teacher evaluation system for months, saying the oversized share of teachers with high ratings illustrated the system was too easy to game and in need of an overhaul" (p. 1).

Although Governor Cuomo and New York State legislature left safety measures in place, teachers are extremely apprehensive by the unsettled control they have over evaluation scores. Valerie Strauss of the *Washington Post* (2015) explained that "Chancellor of The New York State Board of Regents, Merryl Tisch, wants to make new changes" to indicate there is some acknowledgment of the New York State Evaluation system being flawed and not truly obtaining properly scaled numerical values regarding teacher effectiveness. One such example of this flawed system is the case of fourth grade teacher, Dr. Sheri Lederman, who sued the New York State Department of Education for scoring her as "ineffective" in the category of "student growth," which according to the *New York State Education Department 2016/2017 Growth Model for Educator Evaluation Technical Report* (2017), "characterizes the student's current year score relative to other students with similar measured characteristics and prior test score histories" (p. 21). Lederman's students achieved high scores on the math portion of an annual standardized test, but lower scores on the English section. The lower scores on the English portion of the exam caused her to receive a score of 1 out of 20 points, rendering her as "ineffective" for this category.

The New York State Supreme Court in Albany vacated Lederman's low growth score "because of the difficulty in measuring growth for students who already perform above grade level on state tests" (Harris, 2016, para. 6). Lederman had scored 14 out of 20 points in the same category in the prior year. Lederman's attorneys had "elicited affidavits from a number of testing experts…many of whom argued that this and other VAMs were unreliable (i.e., they lacked

consistency over time)" (Close & Amrein-Beardsley, 2018, para. 7)Lederman won her case, with the presiding judge ruling that the state's teacher evaluation system, based primarily on teachers' VAM scores, was "arbitrary and capricious" and "taken without sound basis in reason or regard to the facts" (Close & Amrein-Beardsley, 2018, para. 7). Lederman's case, as well as an outcry from across the teaching profession, was instrumental in state officials voting to exclude test scores from teacher evaluations until 2019. Instances such as this added to the wave of discontent with evaluation systems and have consequently led to revolts by many parents, teachers, and unions.

Teachers across New York State have expressed significant levels of agitation and distrust of Governor Cuomo since he rushed to adopt the evaluation system created by the former education commissioner, John King, which was put into law by the New York Legislature. Tensions between teachers and the state threatened to disrupt a successful transition to a new evaluation system, which created an unhealthy dynamic that threatened to impact students. "Trust [between teachers and administration] facilitates core organizational change processes that instrumentally contribute to improving academic productivity" (Bryk & Schneider, 2002, p. 140). Trust between teachers and the board of education benefits students on multiple levels. Further, trust established between teachers and their districts would allow teachers to focus more on the classroom than job security. What presents even further discontinuity of normed assessments is that, in many schools, "teacher effectiveness often goes unrecognized and poor performance is not addressed" (Weisberg, Sexton, Mulhern & Keeling, 2009, p. 2). Further, Governor Cuomo stated that funding to any districts that did not adopt the evaluation system would stop, which had an immediate impact on the way teachers perceived their value with district administrators. The attempt to reform the evaluation process seems to have conflicting

arguments propelled and orchestrated by individuals who improperly make overwhelming changes to the wrong assessments. "The history of reform efforts in American public education is replete with half-hearted measures, with almost comical misdiagnoses of education problems, with blame-shifting, and with humbug. Everyone is an expert. Most have, of course, suffered through the very system they want to reform" (Hood, 1993, p. 1).

In fact, The Bill and Melinda Gates Foundation expressed concerns about evaluating teachers using a system based on students' test scores prior to teachers being fully acclimated to the standards. In response to that concern, The Gates Foundation released a report of "Initial Findings from the Measures of Effective Teaching Project" (2009). The goal of the MET project was to "improve the quality of information about teaching effectiveness available to education professionals within states and districts—information that will help them build fair and reliable systems for measuring teacher effectiveness that can be used for a variety of purposes, including feedback, development, and continuous improvement" (p. 2). The report included input from 3,000 teachers in six school districts across the nation, with a particular focus on urban districts. According to the report, the Foundation is concerned that the test scores districts are so focused on do not translate into productive feedback teachers can use. Further, the project is concerned that teachers are not getting any feedback aside from scores, which leaves them unable to respond meaningfully. Without feedback, teachers are at a loss for how to improve in the areas that are low-scoring.

However, as of 2018, the data received by The Gates Foundation did not lead to the kind of reform they had hoped. In a report evaluating results from the Gates' program released by the RAND Corporation and the American Institutes for Research (AIR), the MET project did not receive the hoped-for results. The schools participating in the experiment ultimately

> agreed to design new teacher-evaluation systems that incorporated classroom-observation rubrics and a measure of growth in student achievement. They also agreed to offer individualized professional development based on teachers' evaluation results, and to revamp recruitment, hiring, and placement. Schools also implemented new career pathways for effective teachers and awarded teachers with bonuses for good performance. (Will, 2018, para. 5)

However, the schools also reported "there were no big payoffs in terms of improved graduation [rates] or achievement of students in general, and low-income and minority students in particular" (para. 6).

Due to the pushback by teachers and their union counterparts in New York, Governor Cuomo and state legislatures consequently offered a proposition in the 2014-2015 year that would offset the results of any teachers who received "Ineffective" evaluation scores—which was less than 1% of all teachers. "The new system will allow teachers to have their evaluations recalculated without the state test score component for personnel decisions like termination" (Decker, 2014). Moreover, teachers are struggling with the idea of how this method may alter their pedagogy. Teachers have been drawn to listening to various opinions by governmental officials and even trying to differentiate the views held by them, which is often confusing and frustrating as they navigate a bureaucracy that has become increasingly complicated over the years. As such, when "new evaluations are too test-focused, undermined principals, and represented government overreach," an ethical dilemma is created for teachers who have subscribed to older pedagogical principles (Decker, 2014).

As previously outlined, teachers' evaluations do have a place in the educational system; however, the assessment of these implications needs to be fair and equitable. Jacob and Lefgren

(2008) looked at 201 teachers in Grades 2 through 6 and found conclusive evidence that there was a strong relationship between principals' evaluations and value-added ratings (based on student math and reading scores) of the same teachers. Although value-added measures did a slightly better job of predicting future test scores, adding principal ratings increased the accuracy of these predictions. "We find that principals are quite good at identifying those teachers who produce the largest and smallest standardized achievement gains in their schools (i.e., the top and bottom 10%-20%) but have far less ability to distinguish between teachers in the middle of this distribution (i.e., the middle 60%–80%)" (Jacob & Lefgren, 2008, p. 103). The inability to be more precise in identifying teachers in the middle of the distribution, as Jacob and Lefgren described, is at the forefront for why evaluation systems should be corrected. A more precise measurement will bridge gaps between fluctuating statistical data, thereby resolving the problem of inadequate evaluation systems.

Purpose of the Study

Research Questions

The researcher sought to identify the ways in which current teacher evaluation systems are flawed and/or inconsistent in their assessment of teacher skill and, subsequently, teaching quality. Teacher evaluation systems are impacted by several factors that directly impact the school district, individual teacher, and students. Toch and Rothman (2008) would concur, stating:

> a host of factors—a lack of accountability for school performance, staffing practices that strip school systems of incentives to take teacher evaluation seriously, union ambivalence, and public education's practice of using teacher credentials as proxy of teacher quality—have resulted in teacher evaluation systems throughout public education

> that are superficial, capricious, and often don't even directly address the quality of instruction, much less measure students' learning. (p. 1)

Therefore, this study attempted to answer the following questions about teacher critical thinking skills and teacher skill assessment:

- Are there measurable differences in teacher critical thinking skills that have traditionally been evaluated by district administrators?
- Is the rubric that is used and interpreted by administrators accurate, and does it give back a reliable critical thinking score measurement?
- Are the data that justify a rating of "Highly Effective" and "Effective" reliable across two different scales created by the same Pearson Corporation?

Because of link between teacher effectiveness and student college readiness, it is imperative that the results of administrative evaluations be as numerically or categorically reliable as possible with the current APPR rating system in place. This study looked to ascertain if critical thinking scores are aligned using two different assessment tools, both of which were created by Pearson Corporation. A recent press release by Pearson's edTPA (2016), a national assessment for teacher candidates, stated, "More than 27,000 candidate portfolios are included in the findings, and analyses are presented in the report to reaffirm reliability and consistency of scoring, examine evidence of validity and document trends in candidate performance" (p. 1). By assuring accurate evaluation techniques, teachers will be able to transition more effectively into any changes in the education system regarding changes or reforms of teacher evaluations.

Theoretical Perspective

The teaching framework that has been adopted by New York State involves the rubric for features of the APPR scoring rubric and Charlotte Danielson's *Framework for Teaching* (revised

edition 2011). The critical thinking scoring framework will incorporate the RED Model. According to Bennett (2008), "The U.S. Department of Labor (1999) provides the following general guidelines for interpreting a reliability coefficient: above .89 is considered 'excellent,' .80-.89 is 'good,' .70-.79 is considered 'adequate,' and below .70 'may have limited applicability'" (Table 2). Watson-Glaser (2018) offered many additional aspects for appropriate standardization and consistency toward interpretation of the scores that will demonstrate levels of critical thinking abilities. Testing characteristics of the Watson-Glaser test are aligned with the Standards for Educational and Psychological Testing (2014), which indicated:

> Test scores used in psychological assessment ideally are interpreted in light of a number of factors, including the available normative data appropriate to the characteristics of the test taker, indicators of effort, the circumstances of the test taker at the time the test is being given, the temporary stability of the constructs being measured, and the effects of moderator given, the temporary stability of the constructs being measured, and the effects of moderator variables and demographic characteristics on test results. (p. 154)

Such factors are addressed as components of the Watson-Glaser critical thinking assessment test and include: global applicability; business relevance; currency of controversial scenarios and items; equivalence of computer-based and paper-and-pencil forms; background of norms; the most current information on Watson-Glaser II norm groups, including demographic composition found at reliability and standard of error measurement; test-retest reliability; internal consistency; reliability; demographic characteristics to calculate internal consistency coefficients; and content validity.

By proposing such a direct measure of practicality to support teachers' perceptions about the evaluation process, this research will assist in engaging teacher interest in increasing student

achievement levels and college readiness skills by making a more concerted effort to instruct students to "analyze arguments, making inferences using inductive or deductive reasoning, judging or evaluating, and making decisions or solving problems" (Lai, 2011, p. 2). The RAND Corporation has stated that a number of different venues can be used to measure teacher effectiveness—namely student test scores, classroom observation measures, and possible surveys which can be used for feedback on "student engagement and student-teacher relationships. Teaching effectiveness can also be inferred from tests of teachers' knowledge or skills; from teachers' participation in professional development, committees, or mentoring" (p. 1). This statement is aligned with the potential of self-reporting critical thinking scores and how this would significantly compare to districts' APPR evaluation policies. Since the RAND research organization commits to public interest and developing solutions toward policies, suggestions from such an organization would be considered a supplement for application toward present evaluation policies.

Although evaluation systems have been at the forefront of heated debate for their effectiveness for several years, conversations about these systems continue to move the subject forward and garner attempts to continue searching for solutions to constructing effective and accurate teacher evaluations. As outlined in this chapter, the origin of these debates was rooted in the desire of school districts to access funding to serve their communities. The Race to the Top grant started a movement that brought education and professionalism to the forefront of public opinion and engaged communities to prioritize their children's future. While the topic of teacher evaluation is contentious, the ever-increasing focus on education has certainly done much to improve how students are being educated. Ultimately, the goal is for districts to produce students who will be prepared for college and successfully transition to careers that contribute to our

global and local economies. The concept of critical thinking as a skill set is certainly not a new idea and is a valuable tool for both students and teachers to ensure success for both groups.

CHAPTER TWO

REVIEW OF THE LITERATURE

The Evolution of Critical Thinking

This chapter provides a review of relevant research on the teacher assessment category of critical thinking and its direct impact on teaching methodologies. The literature also covers college readiness and its direct correlation to teaching quality, professional development, and impact on student achievement.

While the push to develop critical thinking skills may feel recent with the adoption of the Common Core curriculum, the concept of critical thinking started with Socrates 2,500 years ago. Socrates is credited with establishing the "importance of seeking evidence, importance of seeking evidence, closely" (Found & Hughes, 2016, p. 132). Socrates further asserted that "authority" alone does not constitute absolute knowledge or insight. An investigation with probing question, undergirded with reason and logic, must be performed before an idea can be produced and claimed as knowledge. Plato and Aristotle also believed in critical thinking as a process of systematic thinking, tracing implications that would lead to the revealing of deeper realities. Objectives underlying critical thinking were redirected in the Renaissance between the 15th and 16th centuries. Scholars during these times were influencing higher-level thoughts involving religion, art, society, human nature, and freedom. Inventions and mathematical explanations about planetary movements from such prominent figures as Sir Isaac Newton, Copernicus, Galileo, and Kepler had the purpose of abandoning traditional knowledge, questioning pre-existing theories, and searching for new evidence and sound reasoning. By the 19th century, scientists were increasingly concerned with aspects of human interactions such as social conformity in response to capitalism and questions on how creationism could co-exist with

evolution. Karl Marx and Charles Darwin modernized the use of empirical evidence in the evaluation of social development. In the 20th century, educational reformer and psychologist John Dewey stated, "To educate an engaged citizenry, a prerequisite for a democratic society, schools should teach students how to be problem-solvers to think rather than simply memorize information" (Cam, 2000, p. 160). The preceding historical reflection on critical thinking demonstrates the progression of an ideological exploration spanning several centuries. However, Dewey's later ideas on critical thinking also reflected critical thinking as a fundamental necessity of our current 21st century global society. Education in the 21st century now relies on the instruction of critical thinking to produce students who will be ready to participate in a much more complicated world than what Socrates and his peers experienced. In its study of evolution, critical thinking transitioned from a philosophical to more of a scientific and psychological approach to analyzing thought processes. For this study, the history and refinement of critical thinking, in all of its iterations, justify an examination of how critical thinking skills are assessed as a professional skill set for teachers. In other words, the practical application of Socrates' formula of evidence, analysis, reasoning, and assumptions is tantamount to the mission of this study to evaluate properly how teachers are performing and disseminating information to students.

This study calls for a closer look at how education reforms in teacher evaluations can be optimized for more accurate results. As Hood (1993) stated, "The history of reform efforts in American public education is replete with half-hearted measures, with almost comical misdiagnoses of education problems, with blame-shifting, and with humbug" (p. 1). To protect teachers from erroneous and harmful judgments, multiple and correct measures must be used to

tap evidence of good teaching practices as well as a variety of student outcomes, including but not limited to standardized test score gains.

Critical thinking was at the root of the New York State curriculum overhaul to the controversial Common Core Standards in 2011. In alignment with Common Core Standards, the focus on critical thinking is to ensure that students become proficient in analysis, evaluation, and problem solving. Robin Fogarty, Ph.D. (2012) explained: "The CCSS thread the skills of literacy and reading, writing, speaking and listening through narrative and informative text. The key to implementing the CCSS with relevancy is to address them, with explicit teaching of the higher order thinking skills that are embedded in rich subject matter content" (p. 1). Therefore, tracking a student's ability to use, acquire, and implement these skill sets is largely dependent on the effectiveness of teacher instruction. However, Thomas Angelo (2005) stated that while critical thinking is the "intentional application of rational, higher order thinking skills…students also find these skills difficult to learn, even when provided with direct instruction…. Most college faculty would agree that critical thinking skills are difficult to teach and develop" (p. 6), putting even more pressure on teachers to produce students who already have these skills when they start college. While that may be true for the student experience, the classroom teacher is expected to overcome these obstacles and produce students who are not only proficient in these skills, but who also succeeds with them, depending on the school system and its goals.

According to a recent study by Coggshall, Ott, and Lasagna (2010), most teachers support such a multiple-measures approach as it is the responsibility of teachers to look for the best ways to scaffold children's learning. In addition, it is also the teacher's right and responsibility to question the motives in areas of education that may have an adverse effect on teacher performance. Changes in the New York State teacher evaluation system, specifically,

have drawn adverse attention from both the public and teachers about finding ways to evaluate teacher effectiveness accurately. Daniels (2013) explained how the APPR legislative agenda has changed to align education with notions of "measurement, 'effectiveness,' numerical evaluation, and performance [that] become the sole focus, and the individuals who are deeply involved in and committed to education become byproducts" (p. 27). Unfortunately, the current approach to educational reform epitomized by Education Law §3012-c of 2010 called for more rigorous assessments of teachers and does not follow traditional philosophies of education; rather, it leaves the U.S. education operation to be run by corporations that generate a substantial amount of money. Daniels also stated that funding allocations can be directed by the federal government and further endorsed or manipulated by state officials such as governors who force schools to agree to the Common Core Standards to qualify for Race to the Top cash.

Dewey's "reflective thought" is aligned with evaluating a teacher's own critical thinking levels in comparison to those of an evaluator, such as school administrators. Identifying variables that may lead to a better model of evaluating critical thinking levels would be a significant boon for professional development. VanSickle (2012) wrote that "scientific theoretical analysis creates the possibility of measurement which could lead to more precise, and possible simpler, formulations and tests of a theory" (p. 2). Further, Barry Wadsworth's (2004) "Theory of Cognitive Development" expanded on Jean Piaget's treatment of cognitive development. Wadsworth explained that Piaget's work has "increased our awareness of the egocentric and sociocentric tendencies of human thought and of the special need to develop critical thought which is able to reason within multiple standpoints, and to be raised to the level of 'conscious realization'" (p. 109). All the contributions that were brought from these historical figures have

impacted education in a way that advocates an ideology which aims to maximize critical thought processes.

There has been a progressive change in education over the past century, influences can date back to the industrialization period in the United States. "For example, the transition from small, local economies to larger, industrialized, fast-paced and dehumanizing environments shifted the forms of work that our society practiced as well as valued" (Daniels, 2013, p. 27). Kliebard (2004) drew attention to the social efficiency movement in education, which developed as an educational reform in response to industrialization and the efforts to render the American industry more efficient. Kliebard felt that "It was a science of exact measurement and precise standards in the interest of maintaining a predictable and orderly world" (p. 76). Authors such as Franklin Bobbitt—best known for *The Curriculum* (1918), *How to Make a Curriculum* (1924), and the framework designed by Frederick Winslow Taylor named the "Efficiency of Production" depicted through his book *The Principles of Scientific Management* (1911)—felt that principles of education and curriculum should be reformed and directed more towards scientific measurement. These authors believed that experimental laboratories and schools were discovering accurate methods of measuring and evaluating different types of educational processes. The rationale of study in the book *General Education Exploration in Evaluation* by Dressel and Mayhew (1954) described the possible purpose of evaluation as the "development of more adequate and reliable mean of measurement" (p. 19). The idea that education and the approach to creating curriculum intended toward management, scientific measurement, and precision helped shape the interest of Americans during the industrial period.

Nineteenth century psychologist Lillian Muller Gilbreth was fundamental in creating a bridge between psychology and scientific measurement. The advent of this paradigm shift of

creating a "factory-oriented" view of traditional educational processes is apparent within the terms of the APPR. As Daniels (2013) explained, "This perception of workers results in simplified tasks, increased monitoring of behaviors, and the encouragement of monetary rewards for expected production goals," creating a mindset of suspicion for teachers and how they view the dynamics of quantitative consistency among APPR evaluation scores (p. 27). This insight was supported by Jennifer E. Nauman, the principal at Shields Elementary School in Lewes, Delaware, who stated, "The tendency to be more lenient on a district evaluation is understandable…. Someone's job is in your hands…the rubric is very subjective" (p. 1). In addition to providing a platform for students to have a clear vision for college readiness, "Instructors are urged to provide explicit instruction in critical thinking, to teach the transfer to new context, and use cooperative or collaborative learning methods and constructivist approaches that place students at the center of the learning process" (Lai, 2011, p. 2). Based on the necessity for inter-rater reliability and fair and equitable teacher evaluations scores, this study provided data necessary to align critical thinking skill criteria during the evaluation process. "Policies that create increasingly valid measures of teaching effectiveness—and that create innovative systems for recognizing, developing, and utilizing expert teachers—can ultimately help to create a more effective teaching profession" (Darling-Hammond, 2014, p. iv).

Trends in College Readiness

Education has taken a turn for the worse in the United States, with low college attendance rates, failure to complete college within 4 years, and high college dropout statistics all indicative of the perception gaps that college freshmen and their professors are experiencing. The NCES (2016) stated that "the 6-year graduation rate for first-time, full-time undergraduate students who began their pursuit of a bachelor's degree at a 4-year degree-granting institution in fall 2008 was

60 percent" (para 1). Statistics from the NCES indicated that only 60% of recent high school graduates from a sample of 2,668,496 in the United States who enrolled in college in the fall semester 2008 completed their college 4-year degree by 2014. These numbers are indeed concerning and there may be a need to focus more attention on analyzing freshman college students' perceptions concerning their college readiness and how this differs from the perceptions professors may have about how prepared they are for college-level rigor.

One of the most powerful strategic levers of improvement is to ensure that every student is held to high academic standards. In an environment of high-quality standards, teachers can focus on the higher-order skills that students need to think critically, solve real-world problems, and be successful in the 21st century and beyond. College readiness has been defined as the preparation needed for students to enroll in and subsequently succeed in a postsecondary institution without remediation (Harvey, Slated, Moore, Barnes, & Martinez-Garcia, 2013). Traditionally, students must take a college entrance exam such as the ACT or SAT and graduate high school as well as meet college entrance requirements in order to enroll in a college or university. Due to the rise in students entering college without being prepared for college-level work, some have concluded that earning the high school diploma and meeting college acceptance requirements are not true measures of being college-ready as typically believed (Arnold, Lu, & Armstrong, 2012). A College and Career Readiness student survey administered to 165,000 high school students across the United States found that only 45% of students believed they were ready for college and/or career, yet 87% of students desired to earn a college degree and pursue a career (Leal, 2015). More recently, the American College Testing (ACT) organization reported a decline in the number of ACT-tested high school graduates ready for college based on their test scores. ACT's (2018) report, "The Condition of College and Career Readiness," showed a steady

decline in Math and English since 2014. Further, the report also showed that the percentage of students who met at least three of their ACT College Readiness Benchmarks has also decreased, based on scores collected in 2017 and 2018. The report stated, "a higher number of students this year than in recent years fell to the bottom of the scale, showing little or no readiness for college coursework, [while] thirty-five percent of 2018 graduates met none of the ACT College Readiness Benchmarks" (p. 2). From the postsecondary perspective, university faculty members do not believe that U.S. public high schools are preparing students for the level of work they will experience at a college or university.

In an article written by Schaffhauser (2015) about college readiness for *Campus Technology*, one survey from a series sponsored by a not-for-profit organization called Achieve queried faculty of both 2-year and 4-year institutions and found that instructors believed only 4% of 2-year college students and 12% of 4-year college students were prepared to do college-level work. Additionally, the survey reported that gaps in student readiness for college included math, science, critical thinking, comprehension of complicated materials, development of effective work and study habits, writing and written communication, problem solving, and conducting research. In recent years, the United States has experienced a rise in college enrollment; however, that rise has been accompanied by enrollment gaps along family income lines and dismal college graduation rates (Arnold et al., 2012). Although the number of students applying for college has increased, numbers are dwindling for students who are actually ready to attend college, which then also results in an increased need for college remediation courses to prevent low college completion rates.

Because of these kinds of standardized testing results, the pressure on teachers in the classroom naturally continues to increase. Therefore, the formula for producing a college-ready

student includes measuring the results of student interactions with teachers in the classroom. More than ever, there is a need to have teachers reinforce and foster informative conversations about the level of academic rigor in order to assist high school students with the college transition process. As Conley (2007) stated, "The transition to college has a component of culture shock for students, one that is more severe for students from some communities than others" (p. 5). Without a true sense of college readiness, high school students may acquire a false sense of envisioned reality of key intellectual standings that will be conveyed through college-level courses. Conley added, "Information about the culture of college helps students understand how to interact with professors and peers in college and how to navigate college as a social system and learning environment" (p. 5). Teachers have a large influence on the college readiness of their students (Dunston & Wilkins, 2015; Nagaoka et al., 2013). As a result, examining the perceptions and practices of teachers can provide insight into how teachers influence the college readiness of their students (Springer, Wilson, & Dole, 2014).

Pedagogical Impact on College Readiness: Teacher Effectiveness

High schools have the duty to prepare students for college, especially in these changing politically and economically fragile times. Dunston and Wilkins (2015) asserted that teachers in high school have the responsibility to help students attain grade-level proficiency in reading and math and to be college-ready by the completion of high school (Nagaoka et al., 2013). High school experiences, specifically the rigor of the school's curriculum, have an impact on outcomes for students (Jackson & Kurlaender, 2014). Students should be encouraged to participate in more rigorous courses and be helped to develop better study and time management skills, as these are necessary to prepare students for the rigor of colleges known for their intense curriculum and specialized training in subjects such as Business, Politics, Law and Economics,

Math, Computer Science, and English (Koch, Slate, & Moore, 2012; Springer et al., 2014). This is even truer for students from underfunded and at-risk school districts that have historically had lower high school graduation and college completion rates. Secondary schools also have the important job of socially preparing students who are not considered traditional college material.

Although teachers have the power and responsibility to influence student college readiness (Almager, 2016), teachers who lack critical thinking skills in the application of seeing past racial and socioeconomic constructs can also be a hindrance to college readiness as the teacher is the first real professional interaction students will have. Hence, this is why a focus on professional development is critical to the district's pedagogical success. In the paradigm of teachers, school districts, and teacher evaluations, there needs to be a space to consider the students. Teachers who are highly effective, effective, and developing are at the front lines of college readiness and student educational development because of the push to produce an increase in students who are college-ready as well as successful in high testing scores that align with national levels. The concept of critical thinking skills is deeply embedded in the Common Core curriculum. However, critical thinking skills are also at the forefront of how teachers are rated in their instruction and in the dissemination of curriculum according to the New York State United Teachers (NYSUT) Teacher Practice Rubric of Teaching Standards (2012). One of the goals of this study was to examine and compare reliability between quantitative and qualitative data to find how districts can pinpoint where to make changes in teacher instruction. Reliability of the data is pertinent to professional development in the way that teacher training is more specific.

The pedagogical impact on college readiness has become a significant factor in determining an instructor's efficacy and skill in the classroom. Instructors are compelled to

provide specific instruction toward critical thinking with the prospect that students will learn how to transfer to new contexts while using cooperative or collaborative methods and constructivist practices that place students at the center of the learning process. Teacher effectiveness is defined most simply as a teacher's ability to improve student achievement as measured by student gains on standardized tests (Little, Goe, & Bell 2009). Charlotte Danielson's (2011) Framework for Teaching is a comprehensive evaluation method for teachers used by evaluators, such as administrators. The purpose of proposing this approved teacher practice rubric is to direct teachers toward supplementing the New York State Standards. Teacher success is measured across four domains, as determined by the APPR framework provided by Danielson (2011). See Figure 2.1 for Charlotte Danielson's Framework for Teaching, which is the foundation of the Common Core curriculum.

According to Danielson's Framework (2011), teachers must demonstrate the ability to effectively carry out the requirements of these domains requires the teacher to apply critical thinking skills, as shown in performance indicators within the Framework. They are to also "facilitate students' ability to develop diverse social and cultural perspectives, incorporate perspectives from varied disciplines and use and model interdisciplinary skills in their instruction" (p. 9).

Critical thinking skills are integral to a teacher's classroom instruction and professional development. Within these provisions, this tool looks to encompass a variety of modules to improve teacher performance in light of the new evaluation system that can potentially disenfranchise conventional teachers' pedagogy. The Network for Public Education is also critical of teacher evaluation reforms. In 2016, the organization released a report on findings

from a survey of 2,964 teachers and principals from 48 states on teacher evaluations. The 25-page report covered a range of topics including teacher anxiety, administrator ideas about the

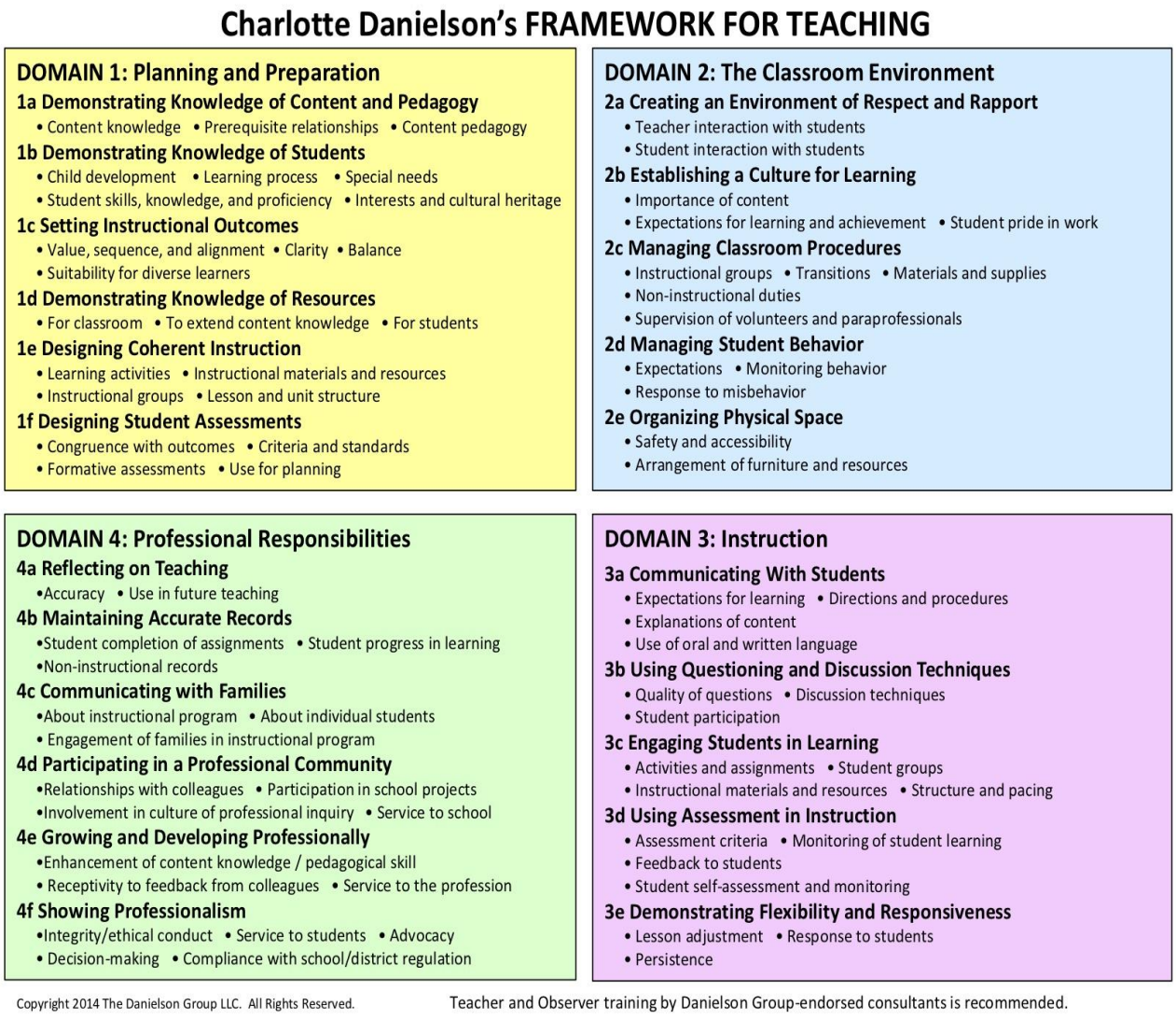


Figure 2.1. Charlotte Danielson's Framework breakdown via domain

evaluation process, and the belief that teacher evaluations "sabotage" teacher development. The report was clear in its criticism of teacher evaluation, going so far as to claim that the evaluation process is in large part responsible for teacher shortages across the country. "When combined with frameworks, rubrics, and high-stake consequences, the nature of teacher evaluation has dramatically changed, and narratives from educators across the United States document that it has changed for the worse" (p. 2). The report further contends: "the emphasis on improving test scores has overwhelmed every aspect of teachers' work, forcing them to spend precious collaborative time poring over student data rather than having conversations about students and instruction" (p. 2).

A coherent objective among teachers toward prosperity in their career is to achieve a profound level of success in the evaluation protocol, which is a mandatory faction toward their tenured-track position. During the process of development, New York State teachers learn and practice district-wide guidelines inherent within each negotiated contract.

Since the onset of the APPR which was signed into law on April 13, 2015, and fast implementation of Common Core student testing, teachers, parents, and students have expressed a sense of controversy about the overwhelming emphasis on student data collecting. Research has indicated that copious amounts of teacher evaluations have led to an ineffective approach to enhancing teacher development: "Policies governing teacher evaluation systems tend to make only vague and weak provisions for professional development, and they fail to ensure that these opportunities are of high quality and of value in improving practice" (Smylie, 2014, p. 97). Furthermore, research has also found evidence that overwhelming teacher evaluations are causing difficulty in designing proper professional development (p. 97). Understanding the differences in teacher perceptions and their ramifications will advance positive future policy

implementations, improvements toward teacher performance, and, conceivably, student achievement. If educators do not see the new evaluation system as a means for improvement, then the system should be adjusted to ensure that best practice is achieved.

Demographic Association

Webb and Thomas (2015) noted that teachers with preconceived expectations of students based on gender, race, and social class could negatively impact student achievement (Bol & Berry, 2005). Another factor to consider is the student demographic with whom teachers are working because not all school districts are created equal. Teachers who are trying to impart the curriculum as per the rubric are also challenged with overcoming factors like the socioeconomic and racial composition of the student body, as well as any students who have cognitive or developmental delays. At-risk school districts are historically underfunded and lack the kind of financial support that a blue-ribbon school district receives. While this may also be regionally connected to urban areas like the five boroughs of New York City, there are districts on Long Island that are in dire need of materials and support for teachers who may be overwhelmed by class size, lack of district support, and students' home support from parents or guardians. Home life, emotional disorders, gang-related activity, homelessness, poverty, and drugs are all factors to consider in general student welfare. Yet while these may be student experiences, what teachers are personally experiencing or how they are affected by their students' circumstances is understandably impactful to teacher effectiveness. Timeliness of certification, training, and/or professional development, as well as the district's access to resources to assist both the teacher and student, are also factors to consider in assessing teacher effectiveness scoring.

Nevertheless, in the case of New York State, the NYSUT Teacher Practice Rubric (2012) expects teachers to "engage students in the development of multi-disciplinary skills such as

communication, collaboration, critical thinking and use of technology" (p. 20) (see Figure 2.2). The NYSUT Teacher's Practice Rubric was approved by the State Education Department as one of five options available to all school districts to meet the new APPR regulations. The Rubric, modeled after the Danielson Framework, dictates that a teacher's ability to think critically is directly tied to the planning and preparation of course material, classroom management, and, finally, instruction, which then places a significant amount of responsibility on teachers to be proficient in how *they* apply critical thinking to their lessons.

***NYSUT's Teacher Practice Rubric* * 2012 Edition ***
Aligned with the New York State Teaching Standards

Element II.2: Teachers understand how to connect concepts across disciplines and engage learners in critical and innovative thinking and collaborative problem-solving related to real world contexts.

NYSED Indicators: *Facilitate students' ability to develop diverse social and cultural perspectives. Incorporate perspectives from varied disciplines and use and model interdisciplinary skills in their instruction. Provide opportunities for students to engage in individual and collaborative critical thinking and problem solving. Teachers model and encourage effective use of interpersonal communication skills to build student capacity for collaboration. Create opportunities for students to apply disciplinary and cross-disciplinary knowledge to personal experiences and real world problems.*

	Indicators	Ineffective	Developing	Effective	Highly Effective
A.	*Incorporates diverse social and cultural perspectives*	Teacher does not plan instruction that facilitates students' ability to develop diverse social and cultural perspectives. Instruction is not aligned with 21st Century skills.	Teacher plans some instruction to facilitate students' ability to develop diverse social and cultural perspectives. Instruction may or may not be aligned with 21st Century skills.	Teacher plans most instruction to facilitate students' ability to develop diverse social and cultural perspectives. Teacher incorporates perspectives from a variety of disciplines and embeds interdisciplinary skills in instruction to align with 21st Century Skills.	Teacher plans all instruction to facilitate students' ability to develop diverse social and cultural perspectives. The perspectives are connected to a sequence of learning both in the discipline and related disciplines and align with 21st Century Skills.
B.	*Incorporates individual and collaborative critical thinking and problem solving*	Teacher does not plan opportunities for students to engage in individual and collaborative critical thinking and problem solving.	Teacher plans occasional opportunities for students to engage in individual and collaborative critical thinking and problem solving.	Teacher plans frequent opportunities for students to engage in individual and collaborative critical thinking and problem solving that align with 21st Century Skills. The teacher models effective interpersonal skills.	Teacher plans on-going opportunities for students to engage in individual and collaborative critical thinking and problem solving that align with 21st Century Skills. The teacher models and encourages effective use of interpersonal skills to build student capacity for collaboration.

90163 Approved by the NYS Education Department – August 2012 8

Figure 2.2. The NYSUT Teacher's Practice Rubric (2012)
This is the model for the APPR test model pertaining to both teacher and student critical thinking objectives.

While this makes sense, one of the questions this study sought to answer was whether there was an accurate way to gauge how much critical thinking is applied to classroom instruction and how to resolve the issue if the level discovered did not meet state and/or district standards.

Researchers have cited a need for additional measures of college readiness apart from the ACT and SAT (Harvey et al., 2013). These additional measures include non-academic factors that influence college readiness. There is agreement in the literature that non-academic factors influence college readiness, but there is discord as to which factors are most influential and should be measured. Possible factors include behavior, motivation, family circumstances, self-efficacy, organization, stress, and time management (Gaertner & Larsen McClarty, 2015; Krumrei-Mancuso, Newton, Kim, & Wilcox, 2013). The factors most frequently measured were self-efficacy, sense of belonging, engagement, and student-to-faculty interaction (Person, Baumgartner, Hallgren, & Santos, 2014). This specific aspect may be relevant to how teachers' perceptions of college readiness are influential in the educational achievements of students. When teachers follow a deficit model towards minority or low-socioeconomic students, it can be detrimental to those students' academic success. For students from low-SES backgrounds, postsecondary education may be unfamiliar because they are not subject to interacting with individuals who have attended college. Teachers have the ability to familiarize these students with how to navigate toward future academic success and help them to cultivate the skills needed to be college-ready (Bok, 2010).

Students' ability to be college-ready is affected by a range of factors not limited to their race and socioeconomic status. There are not many systems to track student progress towards a goal of college readiness; most college readiness standards are based on academic accomplishments and the concept of college readiness is not introduced to students until late in

high school (Gaertner & McClarty, 2015). From 2007 to 2009, there were significant differences in the college readiness rates for Whites, Blacks, and Hispanics, with the greatest difference being between Whites and Blacks (Barnes & Slate, 2013). In almost all years between 1972 and 2008, the college enrollment rates of students from high-income households surpassed the enrollment rates of those from low-income households by at least 20 percentage points (Bernhardt, 2013). Less than 8% of students from low-income households earned a bachelor's degree by the age of 24 (Bernhardt, 2013). An interesting report by the *Lexington Herald-Leader of the Bowling Green Daily News* indicated that those Kentuckians from low-income families who put great pride in school and endeavored among Advance Placement and International Baccalaureate courses were more likely to enroll and succeed in college. The article also stated that Education Commissioner Stephen Pruitt discovered a way to allocate federal funds to assist low-income families in paying for AP testing fees. This resulted in a significant rise from 82,000 students in 1999 to 850,000 students in 2016.

This mechanism came under pressure when Congress and President Obama signed a bill, which placed $28.5 million in testing-fee aid into a block grant. There seems to be a perpetual struggle for some states that rely very heavily on federal funding to cover the cost for financial assistance for AP testing. Since the block funding fell short for the state of Kentucky to help low-income students with AP and IB testing fees, Pruitt reallocated $800,000 in state funds. Another example of college readiness funding came from Acorn Newspapers in an article written by Michael Aushenker (2017) who stated, "Over the next three years, Simi Valley Unified School District will spend about $222,000 in state grant funds on college readiness-related endeavors" (para 1). In 2017, members of the school board along with other trustees unanimously approved its College Readiness Block Grant budget plan, and the guidelines in this plan according to the

California Department of Education will support funds pertaining to programs such as developing advanced-level classes, financing college readiness examinations for students who cannot afford the testing fee, and assisting students with proper counseling services during the process of college admission.

Additional articles have been published on how funding is being directed toward college readiness programs, such as the $500,000 college-ready grant that was designated to help the students of Redlands Unified School District to pursue its goal to better prepare students for college-level tasks. This initiative was coordinated by "a group of local educators, district officials and others are working to establish programs covered by the Department of Education grant, which will serve as a supplement to the Redlands Ready Commitment and must adhere to the state-mandated Local Control and Accountability Plan, a blueprint of sorts for schools," as stated in Hernandez's (2017) article published by the *Redlands Daily Facts Higher Education*. Some guidelines of the Redlands Ready Commitment entails waiving SAT costs for juniors; reducing costs for AP testing; offering opportunities to earn college credit while in high school "as part of a partnership with Crafton Hills College; guaranteed college admission to Cal State San Bernardino and University of Redlands; and college preparation" (para. 7). By engaging students in high school to be a part of classes that are based toward college curriculum, high school teachers have the opportunity to expose high school students successfully to the psychology based around the profound differences between high school and college-level academic behavior while diminishing false pretexts of college expectations (Appleby, 2014).

A significant predictor of whether a child will graduate from college is whether their parents graduated from college (Bernhardt, 2013). Research has shown that there are disparities among students in academic achievement and college readiness when compared by race,

ethnicity, and socioeconomic status (Bernhardt, 2013). One way to challenge this continuous debate is to leave funds available for this group of students, as did the Simi Valley Unified School District in California which left funds available to help socioeconomically disadvantaged students pay for Advanced Placement and International Baccalaureate exams. Aushenker (2017) also wrote, "The state's $200-million College Readiness Block Grant, established by Senate Bill 828, is a one-time grant that California public school districts are receiving to support college readiness among students in grades 9 through 12" (para. 2). Although this group represents one-third of the total student population in California, these data could be used as a recommendation by district leaders and local and state legislatures of all states to open avenues for dialogue on increasing college preparation among disproportionately socioeconomically disadvantaged students. The causes of performance disparities have been debated; however, researchers have hypothesized and found data to support that socioeconomic status, racial and class stereotypes, teacher perceptions, and expectations are influencing factors (Achinstein, Curry, & Ogawa, 2015; Almager, 2016; Webb & Thomas, 2015). The aforementioned sample study was from a district afforded with high economic status, which related to a large amount of parental influence (Asamsama Hemmy et al., 2016) and parental involvement as key indicators of college readiness.

Parental factors such as parents' level of education, parental beliefs regarding student success, and students' perceptions of parental involvement in their academic lives have all been linked to better performance on college readiness tests and an increased likelihood of pursuing higher education (Asamsama et al., 2016). The researchers conducted a cross-sectional study to determine which of these factors related to student academic performance and college readiness. Participants for the study were recruited from a college readiness program conducted in three

high schools in Southern California. The 587 participants and their parents were surveyed and these results were analyzed in conjunction with student grade-level competency scores and Preliminary Scholastic Assessment Test (PSAT) scores. The researchers found a strong relationship between parental expectations and students' academic success; however, the other factors analyzed were not significantly associated with student test scores. Asamsama et al. concluded that while parental involvement is valued and can be a component of student success, the quality of that parental involvement must be taken into consideration as well.

Research has also shown a relationship between student demographic factors and college readiness. Fruchter, Hester, Mokhtar, and Shahn (2012) conducted a study to determine whether students from various neighborhoods in New York City varied in levels of college readiness based on their location. The data used for this study were obtained from the 2011 New York City Department of Education's (NYCDOE) measurement of college readiness indicators from New York City high schools. The data were broken down by zip codes and then by neighborhoods. The strongest association with college readiness was racial and ethnic composition of the neighborhood, with additional factors such as percentage of single mothers, income level, and college readiness scores of students living in that neighborhood. The neighborhoods with the highest percentages of Blacks and Hispanics had the lowest rates of college readiness (Fruchter et al., 2012).

Yet another interesting component of increasing student success in low socioeconomic areas was found through a study by Edmonds (1979), which described that leadership is viewed as especially important in revitalizing failing schools. Relatively speaking, research has found that when there is a strong sense of leadership, there is a high likelihood of student success (Firestone & Riehl, 2005). The researchers concluded that demography is still a leading factor of

academic success as it relates to students' neighborhood. Since this situation still plagues low-socioeconomic areas, we as a nation must continue to incentivize school attendance with quality teachers, increase initiatives for college-ready programs, and improve scholastic opportunities to overcome variables that burden these children residing in such locations.

Many avenues of education and educational testing still need to provide children in low socioeconomic areas with support through educational resources that align fair, valid, and reliable testing conditions that generate high-quality scores. Comparative measures seem to be underrepresented in certain city schools that have inherent problems across racial and ethnic boundaries; an example is "a new analysis by the Office of Comptroller Scott M. Stringer [that] reveals that the graduation gap in city high schools actually widened in recent years" (Stringer, 2016, p. 1). Representation of schools showing increases in graduation rates always sheds light on the result of enhanced policies and changes made to the educational system such as Common Core and new systems of testing and evaluation processes, but fail to emphasize reasons for shortcomings in lower-performing schools that are heavily concentrated with a disproportionate number of Blacks and Hispanics. "The analysis shows that these 110 schools have been on the decline since at least 2010, a downward trend that is largely masked when graduation rates are viewed only from the vantage of the citywide average" (p. 1). This example clearly exposes faults in the educational process that perpetuate across districts, especially in city schools. This issue will ultimately affect college preparation and college readiness. "College readiness rates declined at about 16 percent of schools between 2011 and 2015, with the lowest levels of college readiness clustered in school districts in the Bronx and Brooklyn. Persistent racial gaps exist in college readiness levels as well" (p. 1). Applications of high standards in the educational system need to span demographics while reducing racial and ethnic barriers. In the evaluation of school

districts, teachers and students need to abide by fair, equitable, and reliable systems of educational information.

Since the introduction of the Race to the Top grant in 2009, teacher evaluation systems across the country have undergone critical scrutiny by the public, the media, parents, teachers, and legislation. By 2015, several states have worked to align their standards with college readiness, learning outcomes, and assessments across the country. Because of such a focus on assessment and the negative outcomes feared and experienced, the USDOE revised its requirements and allowed states additional time to adopt new teacher evaluation systems. Subsequently, conversations about teaching and student learning have become more focused on how to meet the needs of both. In New York, opting out of standardized tests is still a movement in which parents are taking the lead in order to protect their children. There is also a better understanding of demographic impact on student learning as children on the cognitive spectrum have become the focus of research that supported the idea that they learn differently but can still be positively affected by mainstream teacher instruction.

More importantly, school districts are finally accepting that student success cannot be solely measured by a singular test. They seem to understand that student growth measurements will fluctuate over the course of a year and do not solely reflect teacher inefficiency. As such, the findings in this study supported the use of multiple measures and student growth to determine teacher effectiveness, but more importantly, they suggested that quantitative analysis is the more likely methodology for comprising a holistic picture of both teaching and learning outcomes. According to Glazerman et al. (2010), the inclusion of both subjective and student data is a step in the right direction since previous research demonstrated that seniority and experience are not appropriate indicators of teacher effectiveness. In addition, this study also maintained that

teacher self-evaluations are more reliable than more subjective assessments. Gravetter and Wallnau (2014) identified several applications for a correlational design, such as making predictions about relationships, demonstrating validity, and evaluating reliability—all of which are needed to understand how student achievement should be integrated into teacher evaluation.

This study, then, aimed to identify the ways in which current teacher evaluation systems are flawed and/or inconsistent in their assessment of teacher skill and, subsequently, teaching quality. Teacher evaluation systems are impacted by several factors that have a direct effect on the school district, individual teachers, and students. The design of the story was comparative in order to identify the ways data collection can be better identified and used for professional development so that teachers are supported in their training and development. The literature reviewed in this chapter covered a range of topics, including the evolution of critical thinking as a concept, its correlation to trends in college readiness, the push to incorporate critical thinking into a Common Core curriculum, its pedagogical impact on student outcomes, and how student learning is impacted by the demographics of the student body.

The next chapter identifies the methodology of this study in its approach to comparing qualitative versus quantitative assessments in order to make the aforementioned identification. Included in the chapter is a breakdown of the Watson-Glaser tools and how they can be used to assess several variables in teacher efficiency, as well as how the Watson-Glaser tools—or tools with similar capabilities—can be used in the future to gain a more holistic perspective of a teacher's skills from a multi-measured perspective.

CHAPTER THREE

RESEARCH DESIGN AND METHODOLOGY

Study Design

The diversified review of the literature presented in Chapter Two outlined the apparent need for a reliable quantitative measure such as the data that can be brought about by the Watson-Glaser Critical Thinking Appraisal. The literature also allocates additional understanding of the importance of critical thinking skills that students should be obtaining from school, how critical thinking skills are being evaluated by district administrators, and how critical thinking skill correspond to college readiness success.

This researcher sought to investigate whether school district ratings are significantly associated with the Watson-Glaser scores and the reliability of these scores in the categories of "Highly Effective" and "Effective," as per the APPR test model NYSUT 2012 3.5b. Teachers' critical thinking evaluation ratings are evaluated by two different rubrics created by the Pearson Corporation in order to assess the reliability of the school evaluator's method to that of a self-reporting critical thinking test method. Therefore, this study was conducted using both the current APPR model and the Watson-Glaser II Critical Thinking Appraisal. Three components were used as part of the critical thinking framework created by Pearson's Watson-Glaser II Critical Thinking Appraisal: recognize assumptions, evaluate arguments, and draw conclusions. These components drove this study in its attempt to evaluate these two different models of assessment so that the results can be useful in aligning consistency in critical thinking evaluation scores. One score was derived from the evaluator's APPR rubric and its value was compared to scores from the self-reporting psychometric online Watson-Glaser II Thinking Appraisal.

Demographic variables were the independent variables: gender, teaching category, years teaching, location raised, and school district ratings. The independent variables of this study were the Watson-Glaser scores.

Research Design

The units of analysis in this study were teachers at multiple grade levels, and the process of measurement proceeded with a quantitative descriptive correlational design. Specifically, the data were ascertained through a survey technique, and information was collected by performing an independent *t*-test. The participants were 74 teachers ($n = 74$) at a Suffolk county public school district on the north shore of Long Island; the majority of teachers at this district were rated "Highly Effective" and "Effective." The district is composed of 778 teachers and 9,405 students. The gender-ratio breakdown of this district was 51% male and 49% female. In terms of demographics, it is also worth mentioning the lack of cultural diversity as compared to the 86% population of White students: 0% American Indian, 1% African American, and 7% Hispanic. The graduation rate for this district in 2016 was 94% in comparison to the National Adjusted Cohort Graduation Rate (ACGR) of 83% in the 2014-2015 school year, as stated by the NCES.

Methodology

This study included participants from a K-12 school district who had been assessed with the APPR testing model. The Smithtown Central School District in Smithtown, New York is comprised of eight elementary schools, three middle schools, two high schools, and 778 teachers. A query was sent to all 778 teachers after the district's superintendent, Dr. James Grossane, granted permission to solicit participation for the study.

Research Questions

- Are there measurable differences in teacher critical thinking skills that have traditionally been evaluated by district administrators?
- Is the rubric that is used and interpreted by administrators accurate, and does it give back a reliable critical thinking score measurement?
- Are the data that justify a rating of "Highly Effective" and "Effective" reliable across two different scales created by the same Pearson Corporation?

Participants

Initial contact with the study participants was via an email soliciting participation by teachers in the district who had previously scored at least an "Effective" or "Highly effective" on the Unannounced Observation or the "Building Administrators Teacher Observational Report" NYSUT 2012 Element 3.5 b section and the Announced Observation which is the "Dept. Admin Independent Evaluator Teacher Observation Report" during the 2016-2017 academic year. Following IRB approval, the researcher sent a bulk email district-wide to subjects eligible to be tested. Participants were given access to the Pearson's Watson-Glaser II Critical Thinking Test computerized link.

Seventy-four of the 778 teachers solicited agreed to take part in the study. The 74 participants had their grades scored, recorded, and sent back to the researcher in an Excel file. Participants were required to answer a survey asking for the following demographic information: name, where they were raised, gender, years of teaching experience, name of school where they were currently teaching, and categorical rating for the prior year on the APPR 3.5b Critical Thinking section.

Instruments

The Watson-Glaser II Critical Thinking Appraisal has normative composition tables, which can be used to rank the scores of the participants (see Figure 3.1 at the end of this section). The teacher's critical thinking scores obtained from this test were compared to normative values and further analyzed in comparing the categorical data (APPR model) and continuous data (Watson-Glaser model) between the two groups of teachers, one representative of "Highly Effective" and the other representing the group of "Effectively" APPR-rated teachers. To obtain quantitative data regarding differences in "critical thinking" skills, the Watson-Glaser II Critical Thinking-Appraisal can be administered either by paper or electronically.

Both distribution models have been found equivalent and raw scores congruent. Each participant group of teachers whose level of critical thinking skills, which was either categorically rated by district administrators as being "High Effective" or lower rank of "Effective," had an overall mean score on the Watson-Glaser II critical thinking appraisal that was compared to a representative normative group set by Pearson. According to Watson-Glaser, "Norms provide a basis for evaluating an individual's score relative to the scores of other individuals who took the same test. Norms allow for the conversion of raw scores to more useful comparative scores, such as percentile ranks" (Watson-Glaser, 2009, p. 14). To assess skills for the purpose of psychometric testing such as teacher's critical thinking levels, it is essential that norms be established that are representative of the general population. In terms of the comparative nature of this type of research, the evidence bases on ranking critical thinking scores can underline difference in skills that will affect instructional capabilities.

The Watson-Glaser II Critical Thinking Appraisal Testing Model. The Watson-Glaser II Critical Thinking Appraisal uses a three-factor model: R (recognize assumptions), E (Evaluate Arguments), D (Draw Conclusions) (RED) (see Figure 3.2 at the end of this section). Pearson has incorporated a confirmatory factor analysis to test the consistency between the constructs of these specific factors. Moreover, the chi square value of 175.66 and 132 illustrates a good overall model fit for this type of investigation. The three-factor model was used to demonstrate, overall, how well the three-factor model explains the critical thinking construct, as the test is intended to assess. Confirmatory factor analysis (CFA) is a beneficial way to ascertain if a testing model can be justified that a relationship exists between the observed variables and their underlying latent constructs. Moreover, CFA is used in social science research and can be useful for the design of the Watson-Glaser II Critical Thinking Appraisal, in addition to aspects regarding interdisciplinary application.

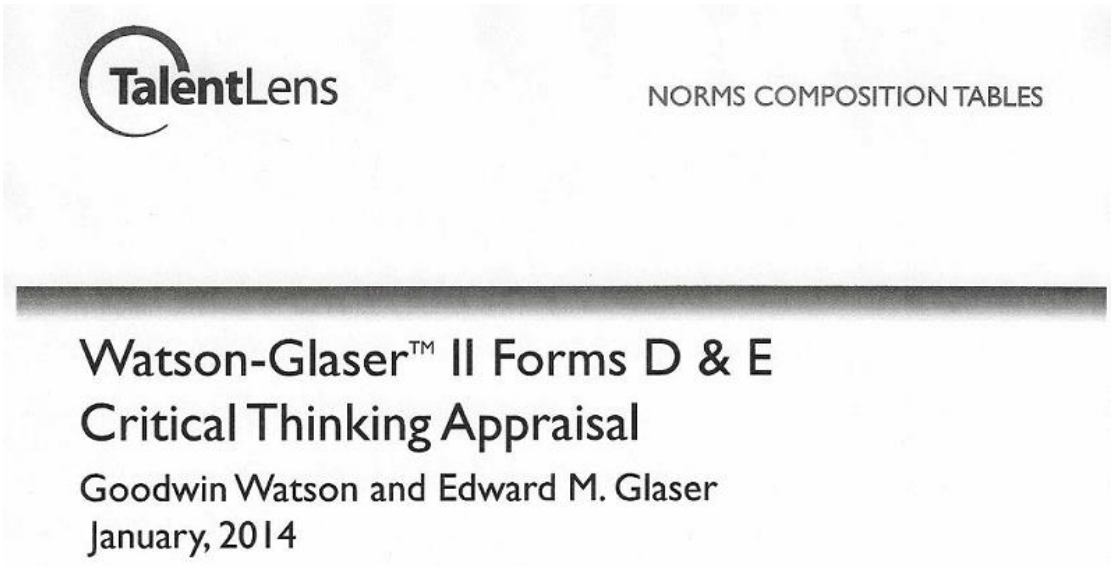
The Watson-Glaser II Critical Thinking Appraisal also indicates a strong linear relationship supporting test-retest reliability showing a correlation coefficient of 0.70 (Watson-Glaser, 2010, p. 20). In addition, CFA illustrates how the Watson-Glaser Critical Thinking Appraisal relates to other measures of cognitive ability, which is an important aspect of construct validity and demonstrates how critical thinking is a unique concept not measured by other tests (p. 25).

Lastly, another reliable quality of the Watson-Glaser Critical Thinking Appraisal is that it works in a similar way to other tests that measure intelligence, like the Wechsler Adult Intelligence Scale (WAIS-IV). The WAIS is commonly used to measure the intelligence or IQ of children and adults alike via a multi-measured approach to assess educational placement and, identify levels of intelligence, learning disabilities, and performance levels, to name a few. This

assessment test is strongly supported by statistical data and attributes a high number of reliable factors that justify strong associations toward defining and testing critical thinking values against the scores obtained by Smithtown's administrator's APPR criteria. If school districts incorporate more precise evaluation techniques along with more reliable standards of quantitatively assigning a rating system that is statistically significant, they will reduce random errors during teacher observations while promoting efficacy consistent across domains in the APPR evaluation criteria.

Watson-Glaser Forms (II Short Form) Scoring Information and Normative Table. The Watson-Glaser Forms II Short Form is scored based on the number of correct items out of the 40 items of which the test is composed. Those raw scores are then converted to a percentile (overall) or standardized (subtests) scores for interpretation relative to a norm group. Scoring levels follow a 30-40-30 (Low, Average, High) percentile range where <30 is considered a low score, a score of 31-70 is represented as an average score, and 71> would be high for the overall critical thinking score. The subtests rely on stanine scores, which is a form of standardized scores where 1-3 is low, 4-6 is average, and 7-9 represents a high score.

Normative composition tables assign normative values to better explore specific results along with providing any user of the experimental tool, the Watson-Glaser Critical Thinking Appraisal, in the case with guidelines for result extremes. Such analysis of data can protect the integrity of the assessment from bias, false interpretations, and generalizations. Normative tables pertaining to a psychometric assessment test like the Watson-Glaser Critical Thinking Appraisal can associate population values and behavioral outcomes to tangible evidence or historical instances. Figure 3.1 indicates normative sample composition tables for occupational norm groups, position type/level norm groups, and educational background norm groups.



Watson-Glaser II – Forms D & E

Norm Group Sample Composition

Occupational Norm Groups

Values represent percentages of the norm group in each demographic category

	Accountant	Consultant	Engineer	Human Resource Professional	Information Technology Professional	Sales Representative
Sample Size	***212***	***264***	***264***	***290***	***480***	***838***
Industry						
Aerospace, Aviation	1	3	2	5	<1	-
Arts, Entertainment, Media	1	<1	<1	1	1	1
Construction	7	2	13	2	1	2
Education	1	5	1	3	5	1
Energy, Utilities	8	5	18	4	3	1
Financial Services, Banking, Insurance	19	-	-	10	16	9
Government, Public Service, Defense	4	7	5	4	3	<1
Health Care	4	7	2	9	10	9
Hospitality, Tourism	1	3	-	3	1	1
Information Technology, High-Tech, Telecommunications	2	14	6	6	42	9
Manufacturing & Production	18	-	31	18	8	11
Natural Resources, Mining	6	1	3	<1	<1	<1
Pharmaceuticals, Biotechnology	1	3	1	4	1	10
Professional, Business Services	8	35	11	6	1	7
Publishing, Printing	1	<1	-	1	<1	5
Real Estate	3	2	-	1	-	1
Retail & Wholesale	6	-	-	8	3	18
Transportation Warehousing	1	2	2	2	2	1
Other/Not Applicable	10	10	5	15	4	14
Position						
Executive	28	18	14	16	11	5
Director	9	8	17	22	22	5
Manager	16	10	27	28	24	28
Professional/Individual Contributor	44	47	38	28	34	30
Supervisor	2	<1	3	1	2	1
Self-Employed/Business Owner	-	11	-	1	1	1
Administrative/Clerical	<1	1	-	2	1	1
Skilled Trades	-	1	1	-	2	-
Customer Service/Retail Sales	-	<1	-	1	<1	15
Other/Not Applicable	1	4	1	2	4	14

Watson-Glaser II – Forms D & E

Norm Group Sample Composition

Occupational Norm Groups

Values represent percentages of the norm group in each demographic category

	Accountant	Consultant	Engineer	Human Resource Professional	Information Technology Professional	Sales Representative
Sample Size	***212***	***264***	***264***	***290***	***480***	***838***
Educational Level						
8th–11th Grade	-	-	-	-	-	<1
HS/GED	-	1	-	2	2	2
1–2 years college	3	3	-	4	9	6
Associate's degree	-	2	-	2	5	2
3–4 years college	2	2	-	5	5	7
Bachelor's degree	63	42	61	47	48	67
Master's degree	32	43	33	37	29	14
Doctorate	-	7	5	3	1	1
Did not report	-	<1	2	1	1	1
Sex						
Female	46	35	10	69	24	34
Male	54	65	88	30	75	66
Did not report	<1	-	2	1	<1	1
Ethnicity						
Asian/Pacific Islander	2	5	6	3	6	4
Black, African American	3	5	2	5	4	4
Hispanic, Latino/a	3	5	6	4	3	4
Native American	1	1	-	<1	1	<1
White non-Hispanic	89	78	82	83	81	84
Other	<1	1	<1	<1	1	<1
Multiracial	-	3	2	<1	2	2
Did not report	1	2	2	3	3	2
Age Group						
16–20	-	1	-	-	<1	<1
21–24	3	4	5	4	5	12
25–29	11	9	9	9	5	10
30–34	13	13	11	11	11	15
35–39	15	11	13	19	16	16
40–49	37	28	31	31	36	31
50–59	19	24	25	19	21	11
60–69	1	8	2	3	2	2
70+	-	-	-	<1	-	-
Did not report	1	3	3	3	2	3

Figure 3.1. Watson-Glaser II Critical Thinking Appraisal normative composition table

Watson-Glaser II – Forms D & E Norm Group Sample Composition Position Type/Level Norm Groups *Values represent percentages of the norm group in each demographic category*	Executive	Director	Manager	Supervisor	Professional/ Individual Contributor	Hourly/Entry-Level	Manager in Manufacturing/ Production
Sample Size	***985***	***828***	***1653***	***163***	***1699***	***357***	***262***
Industry							
Advertising, Marketing, Public Relations	-	-	-	-	-	2	-
Aerospace, Aviation	1	2	1	1	1	1	-
Arts, Entertainment, Media	1	2	1	-	1	1	-
Construction	5	3	6	5	2	2	-
Education	4	4	3	4	8	6	-
Energy, Utilities	2	4	3	12	5	3	-
Financial Services, Banking, Insurance	15	8	9	14	16	21	-
Government, Public Service, Defense	2	4	5	8	6	9	-
Health Care	9	13	9	9	9	4	-
Hospitality, Tourism	6	6	3	1	<1	1	-
Information Technology, High-Tech, Telecommunications	7	8	6	3	10	3	-
Manufacturing & Production	17	20	17	14	8	2	100
Natural Resources, Mining	<1	<1	1	2	2	1	-
Pharmaceuticals, Biotechnology	3	6	7	3	7	3	-
Professional, Business Services	5	3	3	3	9	6	-
Publishing, Printing	<1	1	1	-	1	1	-
Real Estate	2	1	1	1	1	1	-
Retail & Wholesale	7	5	12	9	7	10	-
Transportation Warehousing	1	2	2	3	1		-
Other/Not Applicable	11	10	12	10	8	22	-

Figure 3.1 (continued)

Watson-Glaser II – Forms D & E

Norm Group Sample Composition

Position Type/Level Norm Groups

Values represent percentages of the norm group in each demographic category

	Executive	Director	Manager	Supervisor	Professional/ Individual Contributor	Hourly/Entry-Level	Manager in Manufacturing/ Production
Sample Size	***985***	***828***	***1653***	***163***	***1699***	***357***	***262***
Occupation							
Accountant	6	2	2	3	6	-	5
Adjuster	<1	<1	<1	1	<1	1	-
Administrative Assistant	<1	-	<1	-	<1	20	-
Aircraft Mechanic	-	-	<1	-	-	-	-
Bank Teller	-	-	<1	-	-	1	-
Buyer	-	<1	<1	-	2	-	-
Consultant	5	3	2	1	7	1	-
Customer Service Representative	-	<1	1	6	1	14	1
Distribution and Warehouse Personnel	<1	<1	<1	1	-	-	<1
Engineer	4	5	4	5	6	-	-
Field Service	<1	<1	<1	1	<1	1	1
Financial Analyst	2	2	2	-	6	1	2
General Labor	-	-	<1	2	<1	1	-
Human Resource Professional	5	8	5	1	5	1	6
Information Technology Professional	5	13	7	5	9	2	3
Loan Officer	1	-	<1	-	<1	-	-
Logistics Planner	-	<1	<1	-	<1	1	1
Maintenance	-	-	<1	1	-	<1	1
Marketing Professional	8	12	6	1	4	-	11
Medical Professional	1	2	2	3	3	1	-
Merchandise Planner	<1	-	<1	-	1	-	-
Procurement	<1	1	<1	-	<1	1	2
Product Design	-	1	<1	1	<1	-	-
Production Planning & Scheduling	<1	<1	1	1	<1	-	5
Production Supervisor	-	<1	1	6	<1	-	9
Production Support Personnel	<1	<1	<1	1	<1	-	2
Project Manager	<1	1	1	1	1	1	5
Sales Representative	4	5	14	4	15	8	18
Skilled Trades	-	-	<1	-	-	3	<1
Student	-	-	-	-	-	8	-
Teaching Professional	-	<1	1	-	5	-	-
Underwriter	<1	-	<1	1	1	5	-
Other	57	44	40	55	27	32	27

Watson-Glaser II – Forms D & E

Norm Group Sample Composition

Position Type/Level Norm Groups

Values represent percentages of the norm group in each demographic category

	Executive	Director	Manager	Supervisor	Professional/ Individual Contributor	Hourly/Entry-Level	Manager in Manufacturing/ Production
Sample Size	***985***	***828***	***1653***	***163***	***1699***	***357***	***262***
Educational Level							
8th–11th grade	-	-	<1	-	-	<1	1
HS/GED	2	2	4	13	1	12	6
1–2 years college	4	5	8	15	5	20	6
Associate's degree	1	2	4	6	3	10	2
3–4 years college	2	3	6	12	5	10	4
Bachelor's degree	45	42	51	40	55	42	45
Master's degree	41	42	24	14	25	6	35
Doctorate	5	4	2	1	5	-	1
Did not report	1	<1	1	1	1	<1	<1
Sex							
Female	19	28	34	36	52	69	18
Male	81	72	66	63	47	29	80
Did not report	<1	<1	1	1	1	1	2
Ethnicity							
Asian/Pacific Islander	2	4	4	4	6	4	4
Black, African American	3	2	5	7	6	12	2
Hispanic, Latino/a	4	4	4	4	4	12	4
Native American	<1	<1	<1	1	<1	1	-
White non-Hispanic	89	85	82	82	79	67	87
Other	<1	1	1	2	<1	<1	1
Multiracial	9	1	1	-	2	2	1
Did not report	2	1	2	-	2	2	2
Age Group							
16–20	-	<1	1	-	1	8	-
21–24	<1	1	6	11	11	23	<1
25–29	1	2	6	10	19	17	5
30–34	3	5	13	17	17	14	12
35–39	10	19	18	12	14	9	16
40–49	43	45	34	25	22	15	44
50–59	37	24	17	23	12	12	17
60–69	5	2	2	2	2	3	2
Did not report	1	2	2	1	2	1	3

Figure 3.1 (continued)

Watson-Glaser II – Forms D & E

Norm Group Sample Composition

Educational Background Norm Groups

Values represent percentages of the norm group in each demographic category

	High School Diploma (or GED)	1–2 Years of College	3–4 Years of College	Bachelor's Degree	Graduate Degree (Master's and Doctoral)
Sample Size	***308***	***470***	***322***	***3124***	***2321***
Industry					
Advertising, Marketing, Public Relations	1	<1	2	2	<1
Aerospace, Aviation	3	2	2	3	6
Arts, Entertainment, Media	<1	<1	-	<1	1
Construction	9	7	14	5	2
Education	6	2	4	12	13
Energy, Utilities	2	3	<1	2	1
Financial Services, Banking, Insurance	9	15	10	16	11
Government, Public Service, Defense	3	3	3	6	8
Health Care	9	10	8	8	13
Hospitality, Tourism	3	6	6	3	2
Information Technology, High-Tech, Telecommunications	2	6	5	6	6
Manufacturing & Production	5	5	5	6	8
Natural Resources, Mining	<1	<1	<1	<1	<1
Pharmaceuticals, Biotechnology	-	-	<1	4	7
Professional, Business Services	4	5	3	5	5
Publishing, Printing	-	<1	<1	<1	<1
Real Estate	<1	1	1	<1	<1
Retail & Wholesale	15	15	13	6	4
Transportation Warehousing	3	1	2	<1	<1
Other/Not Applicable	25	15	19	11	10
Position					
Executive	3	7	8	12	21
Director	5	12	7	11	15
Manager	31	33	22	24	22
Professional/Individual Contributor	10	14	19	27	25
Supervisor	5	5	2	2	1
Public Safety	<1	<1	<1	<1	<1
Self-Employed/Business Owner	<1	1	<1	1	1
Administrative/Clerical	10	7	7	3	1
Skilled Trades	2	2	<1	<1	<1
Customer Service/Retail Sales	6	6	8	2	<1
Transportation Trades	-	<1	-	-	-
General Labor	2	<1	<1	<1	-
Other/Not Applicable	24	12	25	17	13

Watson-Glaser II – Forms D & E

Norm Group Sample Composition

Educational Background Norm Groups

Values represent percentages of the norm group in each demographic category

	High School Diploma (or GED)	1–2 Years of College	3–4 Years of College	Bachelor's Degree	Graduate Degree (Master's and Doctoral)
Sample Size	***308***	***470***	***322***	***3124***	***2321***
Occupation					
Accountant	<1	<1	-	5	4
Adjuster	-	<1	<1	<1	<1
Administrative Assistant	4	2	2	1	<1
Aircraft Mechanic	<1	-	-	-	-
Appraiser	-	-	-	<1	-
Bank Teller	-	-	-	<1	-
Buyer	-	-	-	<1	<1
Consultant	3	3	4	4	6
Customer Service Representative	6	8	4	2	<1
Distribution and Warehouse Personnel	3	1	2	<1	<1
Engineer	<1	1	4	5	5
Field Service	1	2	<1	<1	<1
Financial Analyst	<1	<1	<1	4	4
General Labor	<1	<1	<1	<1	<1
Human Resource Professional	2	5	2	6	8
Information Technology Professional	5	7	6	7	7
Loan Officer	<1	1	-	<1	<1
Logistics Planner	-	<1	<1	<1	<1
Maintenance	2	1	<1	<1	<1
Marketing Professional	<1	3	3	5	7
Material Handling Operator	<1	-	-	-	-
Medical Professional	1	2	<1	2	5
Merchandise Planner	-	<1	-	<1	-
New to Workforce	<1	-	<1	<1	<1
Procurement	<1	<1	1	1	2
Product Design	<1	<1	-	<1	<1
Production Assembler	-	-	-	<1	-
Production Machine Operator	-	-	-	<1	-
Production Planning & Scheduling	<1	<1	<1	<1	<1
Production Supervisor	<1	1	<1	<1	<1
Production Support Personnel	<1	<1	<1	<1	<1
Project Manager	3	4	2	5	5
Sales Representative	7	8	8	9	3
Skilled Trades	2	2	<1	<1	<1
Student	17	5	24	14	10
Teaching Professional	-	<1	<1	<1	3
Underwriter	4	3	2	3	<1
Other	30	36	28	23	28

Figure 3.1 (continued)

Watson-Glaser II – Forms D & E

Norm Group Sample Composition

Educational Background Norm Groups

Values represent percentages of the norm group in each demographic category

	High School Diploma (or GED)	1–2 Years of College	3–4 Years of College	Bachelor's Degree	Graduate Degree (Master's and Doctoral)
Sample Size	***308***	***470***	***322***	***3124***	***2321***
Sex					
Female	43	42	46	41	40
Male	56	57	54	58	59
Did not report	<1	1	<1	1	1
Ethnicity					
Asian/Pacific Islander	7	3	7	7	10
Black, African American	6	8	7	7	10
Hispanic, Latino/a	12	10	11	10	9
Native American	<1	<1	<1	<1	<1
White non-Hispanic	71	75	68	71	65
Other	<1	<1	1	1	2
Multiracial	3	2	4	2	2
Did not report	<1	1	1	1	2
Age Group					
16–20	23	4	3	<1	-
21–24	4	3	32	12	2
25–29	3	7	11	18	13
30–34	9	9	7	14	13
35–39	7	14	10	12	15
40–49	31	32	22	26	31
50–59	22	25	13	15	21
60–69	2	5	2	2	4
Did not report	-	-	-	-	-

Figure 3.1 (continued)

Watson-Glaser II Confirmatory Factor Analyses (*N* = 306)

Model	Chi-square	*df*	GFI	AGFI	RMSEA
1	367.16	135	0.85	0.81	0.08
2	175.66	132	0.94	0.92	0.03
3	159.39	125	0.95	0.93	0.03

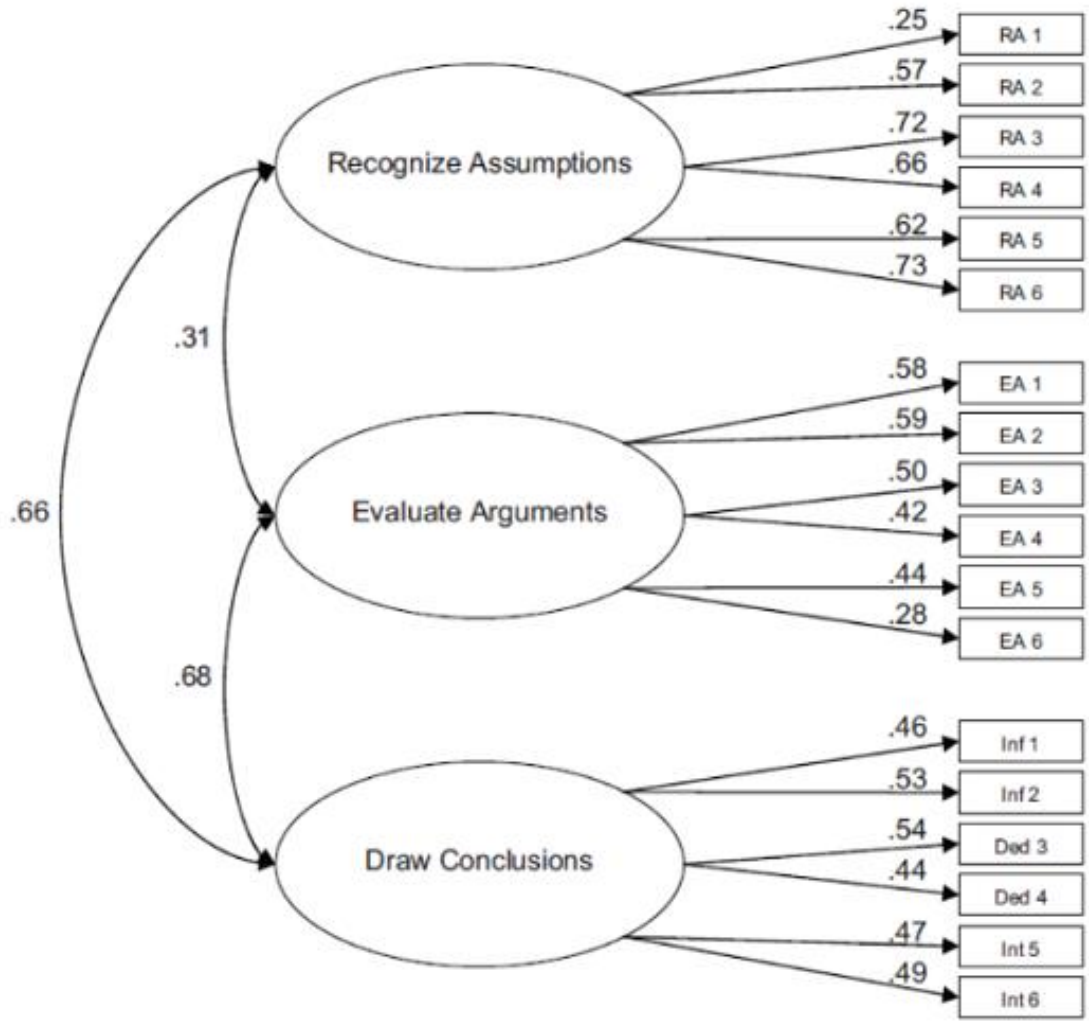


Figure 3.2. Watson-Glaser Three-Factor Model

The following research hypotheses guided the research and were tested in an independent sample t-test analysis.

- H1: The score obtained by the Watson-Glaser II Critical Thinking Appraisal indicates that there is a difference between the means of the Critical Thinking scores of the APPR "Highly Effective" and "Effective" rated schoolteachers.
    - H1a. The Watson-Glaser II Critical Thinking Appraisal indicates that the APPR "Effectively" rated teachers obtained a higher critical thinking mean score than the APPR "Highly Effective" rated teachers. Only the Announced Observation showed statistical significance.
    - H2b. None of the means of the Watson-Glaser scores were significantly higher for teachers rated "Highly Effective." In fact, the mean overall Watson-Glaser score was higher for teachers rated "Effective" than those rated "Highly Effective," which is statistically significant, $p = 0.021$, for the announced evaluations only.
- There is no difference between the means in the level of "Critical Thinking" between "Highly Effective" APPR rated to that of "Effective" rated school teachers with respect to the comparing scores obtained by the school district evaluator's APPR critical thinking score and the score achieved by the Watson-Glaser II Critical Thinking Appraisal.

The first hypothesis, H1, compares average scores on the Watson-Glaser test between two groups of teachers in the categorical district ratings of "Highly Effective" and "Effective." However, the test results in H1a and H2b showed an immediate discrepancy in the mean scores because the teachers in the "Effective" category actually scored higher or had a higher mean score than teachers in the "Highly Effective" categories in both Unannounced and Announced

Observations by district evaluators. One experience is that statistical significance was not reached between the mean scores. The resulting higher scores by the "Effective" rated teachers can be due to chance and is not generalizable. This means that no identifiable causation can be attributed to this kind of result, which is why a multi-measured interdisciplinary approach, in the form of psychometric testing and teacher self-reporting evaluations that are first employed with CFA and chi square to ascertain the goodness of fit of that model, is needed to improve data collection of teacher evaluations accurately and collaboratively. This kind of approach to measurement means districts will be able to identify better why and where a teacher may need improvement in their classroom instruction. This study argues that qualitative data fail to identify data more specifically that can be used for professional development because categorical ratings only demonstrate student success and/or failure rates for which teachers are either lauded or penalized.

The Watson-Glaser 3 Factor Model (Figure 3.2) represents factor analysis as a useful technique for psychological researchers interested in construct validity for scale development, construct validation, or model validation. The relationship among the variables would be for scale development purposes. Thus, the better interpretation of the model would be to demonstrate overall how well the three-factor model explains the critical thinking construct as the test is intended to assess.

CHAPTER FOUR

RESULTS OF THE STUDY

Mean Score Comparisons of Significant Values

Inferential statistics were used to verify the reliability of Smithtown's school district administrator's categorical rating method of evaluating teachers' critical thinking skills comparable to the raw score values obtained through the "self-reporting" Watson-Glaser II Critical Thinking Appraisal. To properly assess the reliability of scores for a teacher's critical thinking level by comparing values obtained from the categorical rating of school district evaluators and values acquired from the raw score results of the Watson-Glaser II Critical Thinking Appraisal, t-tests were used. This statistical method compared the averages of the continuous data items in the Watson-Glaser scores by two-level categorical data elements, which included the school district ratings. It was found that a total sample of teachers' (N = 74) critical thinking skills during an "Unannounced" observation were rated "Effective" by their school district administrators. A sample of these teachers (n = 24) had a higher mean score of 22.33 on the Watson-Glaser II Critical Thinking Appraisal. The other sample of 50 teachers' critical thinking skills was rated "Highly Effective." This particular group, however, had a lower score of 20.22 on the Watson-Glaser II Critical Thinking Appraisal. Although the difference between the means of the Watson-Glaser II Critical Thinking Appraisal scores for the Unannounced sample did not reach statistical significance, it is of note that the "Effective" rated teachers had a higher average score on the Watson-Glaser II Critical Thinking Appraisal (M = 23.33) than did the 50 "Highly Effective" rated teachers who had an average score (M = 20.22) (see Table 4.1).

After computing the means of the two samples, a t-test was performed to establish if there is an actual difference between the two groups. The Levene Test for Equality of Variances

(Table 4.2) was used to determine if the variances of the two sets of sample teachers were similar. Further analysis of Table 4.2 verified that the *p* value of 0.284 was larger than the alpha value (α) of 0.05, indicating that the null hypothesis was accepted and the variances of the scores were not significantly different; hence, the equal variance t-test can be used. The "Sig (2-tailed)" table column showed a *p* value of 0.123; this again is higher than an alpha value (α) of 0.05, implying that even though the sample of "Effective" rated teachers had a higher average Watson-Glaser II Critical Thinking score, we cannot claim that the average scores are actually different statistically. In addition, Table 4.2 indicates confidence intervals between -.589 and 4.816, which revealed that 95% of the time the difference in scores will be between -.589 and 4.816. However, because no other statistical comparisons can be made from this specific part of the research, a larger discussion begs the question of what type of sound, reliable, and statistically significant evaluation models are being accepted by school boards and implemented by state legislators.

Table 4.1. *Means and Standard Deviations for a t-test (SPSS Output)*

Null Hypothesis: There is a difference between the means scores of the Watson-Glaser Critical Thinking Appraisal of each group of teachers, "Effective" district-rated and "Highly" Effective district-rated. Rejecting the Null or Alternate Hypothesis: There is no difference between mean scores.
Levene's Test Results • If $p < 0.05$, reject null hypothesis and accept alternate hypothesis. The variances are significantly different, so assume they are not equal. • If $p > 0.05$, accept the null hypothesis. The variances are not significantly different, so assume they are equal.
T-test Results • If $p < 0.05$, reject null hypothesis and accept alternate hypothesis. The means are significantly different. • If $p > 0.05$, accept the null hypothesis. The means are not significantly different.

Means and Standard deviation for a t-test (SPSS output)
Group Statistics for the Unannounced Observations

Table 4.2. *Independent Sample Test*

	Rating Unannounced Rev	N	Mean	Std. Deviation	Std. Error Mean
RA	1.00 Effective	24	6.29	3.210	.655
	2.00 Highly Effective	50	4.92	2.842	.402
EA	1.00 Effective	24	6.96	1.601	.327
	2.00 Highly Effective	50	6.98	1.890	.267
DCInferDeduce InterpretItems DC (Infer+Deduce+ Interpret Items)	1.00 Effective	24	9.08	3.335	.681
	2.00 Highly Effective	50	8.32	2.591	.366
Overall	1.00 Effective	24	22.33	5.858	1.196
	2.00 Highly Effective	50	20.22	5.262	.744
GradDegreePctle	1.00 Effective	24	22.88	25.149	5.133
	2.00 Highly Effective	50	15.40	19.398	2.743

Table 4.2 (continued)

		Levene's Test for Equity of Variances		t-test for Equity of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
		Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	of the Difference	
									Lower	Upper
RA	Equal variances assumed	1.660	.202	1.863	72	.066	1.372	.736	-.096	2.839
	Equal variances not assumed			1.785	40.848	.082	1.372	.769	-.181	2.924
EA	Equal variances assumed	.789	.377	-.048	72	.962	-.022	.448	-.914	.871
	Equal variances not assumed			-.051	52.934	.959	-.022	.422	-.868	.825
DCInferDeduceInterpretItems DC (Infer+Deduce+Interpret Items)	Equal variances assumed	3.183	.079	1.079	72	.284	.763	.708	-.647	2.174
	Equal variances not assumed			.987	36.804	.330	.763	.773	-.803	2.330
Overall	Equal variances assumed	1.163	.284	1.559	72	.123	2.113	1.356	-.589	4.816
	Equal variances not assumed			1.500	41.350	.141	2.113	1.408	-.730	4.957
GradDegree Pctle	Equal variances assumed	5.138	.026	1.406	72	.164	7.475	5.315	-3.121	18.071
	Equal variances not assumed			1.284	36.611	.207	7.475	5.821	-4.323	19.273

There are three scales in the Watson-Glaser II Critical Thinking Appraisal data: recognize assumptions (RA), evaluate arguments (EA), and draw conclusions (DC), the last being a combination of three subscales. The table above indicates the results comparing the mean score values of the Watson-Glaser Critical Thinking Appraisal, grouped by the district administrator's "Unannounced" observation and the district administrator's categorical rating as per the criteria with the APPR section 3.5b "critical thinking" section. The average overall score for those rated Effective, 22.33, was higher than the average for those rated Highly Effective, 20.22 ($p = 0.123$). Although a *p*-value of 0.123 is not statistically significant, this analysis still indicated results that were counterintuitive to the way district administrators conduct their analysis on critical thinking levels during teacher observations, aligning with this study's research question. Those rated "effective" had a higher Watson-Glaser score average than those rated "highly effective," which may indicate a problem with the way the district evaluation is done. Although the margin between the two results is narrow, the results from the district administrator's evaluation have a significant impact on teacher development and employment. Identifying this discrepancy is one step in a multi-step process to fine-tune how a teacher's efficacy is evaluated across all other evaluation criteria.

The *p*-values are indicated in Table 4.3; however, normally statistically significant results are mentioned in the analysis summary, except to mention that the others are not significant, i.e., >0.05. On the announced observation for section 3.5b of the New York State APPR school district's administrative criteria for assessing a teacher's level of critical thinking skills, the results were contradictory as it was found that the average overall score on the Watson-Glaser II Critical Thinking Appraisal for those rated Effective, 22.96, was significantly higher than the average for those rated Highly Effective, 19.86 ($p = 0.021$). It is evident that these results were

also counterintuitive, which may indicate a problem with the way the district evaluation is done. The average overall score for those rated Effective, 22.96, was significantly higher than the average for those rated Highly Effective, 19.86 ($p = 0.021$). In this specific scenario of the research, the *p*-value reached a statistically significant level, indicating results showing that the differences between the mean scores of the two groups of teachers were not likely to have occurred by chance. This finding is important for academic disciplines and practitioners that rely comprehensively on analyzing data and research. While this particular research sample was a small pool of only 74 participants, it is reflective of the concern this study addressed in the alignment of teacher evaluation. The results of this study did take into account that scores may be different with a larger scale test pool; however, the researcher believes the results reflect the necessity for a reform process to gauge evaluation assessments correctly for teachers in New York State.

The independent sample t-test for the announced observations indicated that 25 teachers who were rated by district administrators as having an "Effective" critical thinking value had a higher mean score on the Watson-Glaser II Critical Thinking Appraisal than did the 49 teachers who were rated "Highly Effective" by the district administrators. The Announced observation data pertaining to the mean score on the Watson-Glaser II Critical Thinking Appraisal did, in fact, indicate statistical significance; the mean for the "Effective" rated teachers was higher than that of the district-rated "Highly Effective" teachers. The "Effective" rated teachers had an average mean score ($M = 22.96$) and a standard deviation of 5.81, while the data for the "Highly Effective" rated teachers had an average score ($M = 19.86$) and a standard deviation of 5.10, as referenced in Table 4.3. Levene's test was performed to ascertain if the variances were equal among the two groups of teachers. The results in Table 4.4 verified that the *p*-value of 0.107 was

Table 4.3. *Means and Standard Deviations for a t-test (SPSS Output): Descriptive Statistics for the Announced Observations*

	Rating Announced Rev	N	Mean	Std. Deviation	Std. Error Mean
RA	1.00 Effective	25	6.28	3.195	.639
	2.00 Highly Effective	49	4.90	2.838	.405
EA	1.00 Effective	25	7.12	1.424	.285
	2.00 Highly Effective	49	6.90	1.960	.280
DCInferDeduce InterpretItems DC (Infer+Deduce+ Interpret Items)	1.00 Effective	25	9.56	3.380	.676
	2.00 Highly Effective	49	8.06	2.427	.347
Overall	1.00 Effective	25	22.96	5.813	1.163
	2.00 Highly Effective	49	19.86	5.099	.728
GradDegreePctle	1.00 Effective	25	24.96	25.077	5.015
	2.00 Highly Effective	49	14.18	18.750	2.679

Table 4.4. *Levene's Test for Equality of Variance in a t-test (SPSS OUTPUT)*

		Levene's Test for Equity of Variances		t-test for Equity of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
		Equality of Variances		t-test for Equality of Means						
									of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
RA	Equal variances assumed	2.112	.150	1.899	72	.062	1.382	.728	-.069	2.833
	Equal variances not assumed			1.826	43.664	.075	1.382	.757	-.143	2.908
EA	Equal variances assumed	2.929	.091	.502	72	.617	.222	.442	-.660	1.104
	Equal variances not assumed			.556	63.288	.580	.222	.399	-.576	1.020
DCInferDeduceInterpretItems DC (Infer+Deduce+Interpret Items)	Equal variances assumed	6.009	.017	2.192	72	.032	1.499	.684	.136	2.862
	Equal variances not assumed			1.973	37.012	.056	1.499	.760	-.041	3.038
Overall	Equal variances assumed	2.668	.107	2.361	72	.021	3.103	1.314	.483	5.723
	Equal variances not assumed			2.262	43.213	.029	3.103	1.372	.336	5.869
GradDegree Pctle	Equal variances assumed	7.725	.007	2.081	72	.041	10.776	5.179	.452	21.100
	Equal variances not assumed			1.895	38.094	.066	10.776	5.686	-.733	22.286

The means of the Watson-Glaser Appraisal subscores—Evaluate and Recognize; the combined subscores—Infer, Deduce, and Interpret; as well as the total or overall score and graduate percentile were compared using t-tests to determine if the means were significantly different for Effective and Highly Effective results on the school district evaluations. The results illustrating the comparison of the "Unannounced" evaluations by district administrators did show a difference in terms of the mean score of the overall Watson-Glaser II Critical Thinking Appraisal score and that of the assigned categorical rating of the school district (e.g., $n = 24$ effective rated teachers scored a 22.33); this was higher than those teachers who were categorically rated as "Highly Effective" (e.g., $n = 50$ effective rated teachers scored a 20.22), as illustrated in Table 4.1.

The results depicting comparisons of teachers during their "Announced" evaluations by district administrators also demonstrated counterintuitive results in addition to showing statistical significance. The average overall mean score for the Watson-Glaser II Critical Thinking Appraisal of a sample (e.g., $n = 25$ teachers rated effective was 22.96) was significantly higher than the average score for the sample (e.g., $n = 49$) "Highly Effective" district-rated teachers, which had a mean average of 19.86, as illustrated in Table 4.2. Furthermore, the subsequent Independent Samples Test of the "Announced" evaluation t-test for equality of means was statistically significant ($p = 0.021$).

Announced Evaluations Results

1. The average overall score for those rated Effective, 22.96, was significantly higher than the average for those rated Highly Effective, 19.86 ($p = 0.021$).
2. Note: For the combined score (Infer, Deduce, Interpret) and graduate percentile, Levene's Test was significant, meaning that the variances of the groups were

different; hence, the unequal variance t-test needs to be used. The p-values for the unequal variance *t*-tests were >0.05, i.e., not significant.

3. There were no significant differences on the other Watson-Glaser measures.

Demographic Results

Analysis of variance was used to compare the means of the Watson-Glaser II Critical Thinking Appraisal scores by demographic measures. The means of RA, EA, DC (Infer, Deduce, and Interpret), Overall and Grade percentile did not significantly differ by location raised, teacher type, gender, or years teaching.

Summary

The average Watson-Glaser scores significantly differed in only one instance. The overall mean score for the Watson-Glaser II Critical Thinking Appraisal was significantly higher for those rated Effective on the announced observation, compared to those rated Highly Effective, which is contrary to expectations from the categorical ratings proposed by the district administrators. Although not statistically significant, the overall mean score for the Watson-Glaser II Critical Thinking Appraisal was higher for the Effective Group compared to the Highly Effective group for the unannounced observation, which again was an inconsistent result. The lack of significant demographic associations in the data confirmed that no bias existed that could possibly have influenced the results of the school ratings versus the Watson-Glaser scores. The lack of racial and cultural diversity in the sample suggested there could be no prejudicial variance in how test subjects answered the questions.

CHAPTER FIVE

SUMMARY, CONCLUSIONS, IMPLICATIONS, AND FUTURE STUDY

Interdisciplinary Contributions

This study drew from several professional frameworks across the fields of psychology, social science, economics, technology, statistics, and educational historiography in order to imagine a more accurate evaluation system in the field of education. The scaffolding of this knowledge took a holistic approach to improving teacher evaluation systems and, by virtue, professional development. Therefore, this kind of interdisciplinary inquiry can only improve evaluative statistical data, as the inclusive data provided several viewpoints from which to analyze and apply statistical outcomes. Further, an interdisciplinary approach ultimately enriched this research and its structured analysis of the objectives involved with properly analyzing critical thinking scores with enhanced and widely used statistical models such as the Statistical Package for Social Science (SPSS).

The comparisons of evaluation systems made throughout this research were used for the purpose of conducting a thorough analysis of a statistical program like the SPSS because it reflected a popular method that researchers use in such fields as psychology and sociology. The original SPSS manual (Nie, Bent, & Hull, 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis. Another area of this research that illustrated and established the need for a reciprocal relationship was the method by which the teachers' critical thinking skills were ascertained. The three-factor model of the Watson-Glaser II Critical Thinking Appraisal was tested by the Pearson Corporation by using confirmatory factor analysis, which is also used by social science researchers. The Watson-Glaser II Critical Thinking Appraisal is a widely used psychometric test for determining and

evaluating managers in organizations of all types and venues. This test also has international influence and can be used uniformly in standardization in terms of reliability and validity, making it a sound and trusted testing tool for use in psychological testing. As such, the framework of this study is applicable to other school districts and, therefore, can be proposed as a universal framework toward analyzing and verifying APPR scores.

The statistical data analyzed in this study, from comparisons of the Watson-Glaser II Critical Thinking Appraisal to the current evaluation system, joins similar scholarship that values the use of an interdisciplinary lens to assess current trends in education from both administrative and pedagogical perspectives. Allen F. Repko's (2014) work, *Interdisciplinary Research: Process and Theory*, focused on the benefits of interdisciplinary teaching. Repko insisted that interdisciplinary classes promote "perspective taking and thinking critically about conflicting information on an issue or problem from multiple knowledge sources" (Repko, Szostak, & Buchberger, 2014, p. xviii). While Repko's work centered on teaching, his approach to education works from both sides.

Repko's work is an example of burgeoning trends in educational perspectives that have significantly advanced since the advent of the age of industrialization. In the 19th and 20th centuries, there was a belief that education should aim its sights on preparing children for productivity. The idea of preparing children for work instead of looking at the child's innate nature to explore and develop overall more stresses views about limiting waste in the educational process. We now know that efficiency and production come from critical thinking skills as part of a multilevel education that teaches children to anticipate other social situations outside of "work." Moreover, theories by Kliebard (2004) based education on the idea that there must be a humanistic approach associated with the component of following a "social efficiency model,"

which today is threatened by educational reforms embedded within APPR. Along with humanism's hold on America's curriculum, an accountability-based testing regime has arisen.

Measuring student performance in traditional humanist curricular areas necessitates that "schools devote substantial resources to English, Math, Social Studies, and Science classrooms" (Dake, 2011, p. 208). Furthermore, this research will help defend and preserve authentic instruction by educators and contribute to identifying factors in current systems of evaluations that may not be aligned quantitatively. It will also be valuable for districts to create a framework that will enforce coherence in assessing a teacher's performance in the area of critical thinking while exposing ineffective APPR scoring methods. By having the use of programs that can collect and analyze data to verify APPR evaluation scores, the creative partnership between computers and the psychometric assessment used in this study presents an interdisciplinary interconnectedness that helps teachers to recognize and identify the positive contributions of such research. New technology offered to researchers affords the ability to test large sample sizes expeditiously. With the advent of computers, there are now means to send district-wide email notifications about this specific investigation. Computers give access to specialized website links that can collect and interpret correspondents' responses to surveys and questionnaires. Programs such as Google documents can assist with organization and editing tools and offer the ability to share information to other faculty. This specifically provides the researcher with the ability to notify participants on how to access the Watson-Glaser Critical Thinking Appraisal.

The role of political influence, which also illustrates the contributions of social science to academic disciplines while serving standard implementation processes, is affected from time to time. An idealistic evaluation system put in place to satisfy political agendas may have different implications for the vast number of districts across New York State. The prospects of this

research may bring alternative and substantive value to methods that may, in fact, prove to enhance critical thinking evaluation standards, while also devoting attention to instilling strategies to increase student achievement levels. It will be important that political figures and local members, such as school board members and superintendents, come together to help decide what is the best proposition for increasing reliability among evaluation systems, which is predicated upon teachers' APPR scores, "…collaborations across disciplines, integration of past disciplinary efforts to create a new perspective, and the synergy created by central facilities that bring people together" (Pellmar & Eisenberg, 2000, (p. 1). During each election cycle, depending on the political viewpoints of legislative figures elected into governmental offices such as the U.S. Department of Education and even as high as the Executive Office of the President (EOP), changes will be imposed or recommended in policy initiatives, review pre-existing laws pertaining to education will be reviewed, and research data will be reassessed and re-evaluated.

This research looked into ascertaining balance and reliability in teacher evaluation systems created as a result of a political agenda. Although initiating a plan to approve college- and career-ready standards is advantageous for education, there still may be unsettling commitments between federal and state control over the application and implementation of programs that districts must adopt and bear responsibility for along with the burden of its effectiveness. Researching more sound methods of evaluating and aligning teachers' critical thinking skills may, in turn, lead to a more prosperous view of educating students. The ultimate goal will be student-centered, and if districts are enhancing evaluation methods in areas such as "Critical Thinking," then students will potentially experience more growth in areas such as "Recognizing Assumptions," "Evaluating Arguments," and "Drawing Conclusions" about college-level work in addition to being successful in the workplace.

One method that Pearson uses to test each construct in the three-factor model (RED) is by performing confirmatory factor analysis. This type of statistical model is usually operational in social science. Contributions in statistics were used in this study in various forms: the Watson-Glaser II Critical Thinking Appraisal, Test-Retest Reliability, and Internal Consistency Coefficients between Demographic Characteristics. This assessment related to other measures of cognitive ability by Convergent Validity, while the Critical Thinking Appraisal related to important work outcomes such as job performance by reporting Descriptive Statistics and Correlations.

In terms of the interdisciplinary contributions made in the field of psychology, Pearson's critical thinking self-reporting evaluation scale can be used for clinical and psychometric measures. This innovative self-reporting test can evaluate specific objectives inquired about by employers while acting as a helpful supplement to improve job performance—in this case, fostering higher critical thinking skills in students. Psychometric assessment can be a valuable tool in the workplace, especially with teachers, whereby progress in student achievement relies heavily on cognitive abilities. Psychometric testing is now used by over 80% of Fortune 500 companies in the United States. These types of assessment tests can help evaluators find the most suitable working environments for new employers, in addition to providing management with guidance on career progression for existing employees.

Measuring an individual's cognitive ability via self-reporting psychometric assessment rather than relying on an evaluator's classroom observation proves to be more valuable and offers evaluators specific and reliable quantitative data on a teacher's actual critical thinking evaluation score. Beside teacher certification exams, more comprehensive data can be obtained from reliable psychometric and statistically significant tools such as the Watson-Glaser Critical

Thinking Appraisal. Quantitative assessments such as Pearson's Critical Thinking Appraisal can have additional functions in school districts, such as determining the effectiveness of not only comparing administrators' evaluator scores but also portraying the quality of professional development programs.

As mentioned earlier, much educational development in New York State is affected by economics, i.e., funding allocations voted by the state legislation, county and town local governments, and finally individual school district budget votes. School districts across New York State did not anticipate the economic burden created by the implementation of APPR and the massive high-stakes testing. Although federal funding was offered under the assumption that new evaluation systems would be followed, part of the objectives behind researching more advanced evaluation tools is to foster interdisciplinary education in economics to develop sustainable economic policies in school districts. School boards must be prepared to extend beyond economic restrictions if, in fact, more precise quantitative assessments have been tested and deliver better data. If this type of research and self-reporting critical thinking assessments are used in school districts, there would be a need to finance testing fees with additional revenue costs within the school budget, but the positive result of more accurate evaluation scores will outweigh the cost. The economic soundness of voters will determine passing any increase in budgetary expenses, such as the cost for purchasing Pearson's Watson-Glaser Critical Thinking Appraisal or even providing additional funds for professional development in critical thinking courses.

By exploring different methods of improving components within the APPR evaluation system, more teachers will support leadership roles in encouraging critical thinking strategies in their students. Duerr (2008) of "Interdisciplinary Instruction" explained the importance of

broadness to students' futures in how "their cognitive development allows them to see relationships among content areas and understand principles that cross curricular lines. Their psychosocial development gives them the ability to understand people and to look at situations from various viewpoints" (p. 177). By researching areas in education that supply teachers with a better platform for evaluation methods, students will benefit from engaging in more advanced-level critical thinking applications. This is aligned with Hilary Staples (as cited in Jones, 2009), who is an AP consultant for the College Board at San Domenico School and also noted that the integration of interdisciplinary studies offers students "advanced thinking skills leading to discovery and real-world problem solving" (p. 16). In closing, Jones (2009) stated that "Students and their teachers will advance in critical thinking, communication, creativity, pedagogy, and essential academia with the use of interdisciplinary techniques" (p. 80).

Benefits and Contributions of the Study

One of the implications of this study is whether teachers who scored "Highly Effective" on their APPR are using a greater degree of "critical thinking" in preparing high school senior students for college and if they are doing so more effectively than teachers who scored "Effective." If, however, it is found that "Effective" APPR-rated teachers have a higher level of critical thinking skills as per results of the Watson-Glaser Critical Thinking Appraisal, especially in the area of the Pearson's APPR NYSUT 2012 3.5 b, then school evaluators need to re-evaluate the rubric they are using and align the observation process to represent more accurately the teachers' critical thinking skills.

The data collected from this study can contribute to the scholastic community by offering a platform for an evolving set of presentations, professional development, and course content designed to support teachers to encourage increasing college readiness preparation and critical

thinking skills while obtaining higher APPR scores on their evaluations. Districts statewide can choose the most qualifying professional development programs designed for each district's demographics and budget. This study contributes to positive social change that applies to teachers and evaluators in attaining more precision in APPR scoring, specifically within critical thinking criteria. Two interesting statistical facts parallel the importance of the possible results of critical thinking skills: high college level achievements and college graduation rates; these data were published in *The Long Island Index.* Another statistical trend is a reduction in population growth on Long Island, from 267% between the years of 1930 through 1970 to only 11% from 1970 through 2010, along with a steady decrease in the average annual employment growth rate. According to the U.S. National Bureau of Economic Research, the U.S. recession began in December 2007 and, within 18 months, many manufacturing companies—in particular, the defense industry—lost contracts and were forced to close engineering plants, resulting in the loss of thousands of jobs. Long Island has lost its competitive edge in the employment growth it once enjoyed. According to a report by *The Long Island Index*, "Our population of residents aged 18 to 34 has declined steadily over the past four decades, from more than 16% of the total to about 10% today. That represents a loss of 150,000 future leaders and a frightening drop in the economic vitality of the region" (p. 1). A significant benefit of aligning teacher evaluation scores, especially in the area of critical thinking skills, is to improve college readiness, which, in turn, increases college graduation rates. A younger population of higher-performing college graduates will add prosperity to the working force.

Discussion

The most important aspect to consider in assessing the practical use of the New York State teacher evaluation system, is the consideration of its ability to deliver statistically reliable

results. In addition, the district must also consider and be flexible, in accommodating teachers' instructional practices, provide adequate training in teacher observations, and instill legitimate guidelines consistent in using equitable measures across various districts and disciplines. School boards must take the responsibility for adopting a uniform evaluation system that provides teachers with an overall rating score that reports valid, accurate, and reliable measures demonstrating high-quality statistical data. The variances in mean results in the Announced and Unannounced Observations alone suggest that categorical ratings are arbitrary and do not accurately reflect a teacher's efficacy as espoused by student scores.

The research questions posed in Chapter One of this study reflected on measurements of teacher efficacy and its connection to test scores. The intent of the study was to ascertain not only how to identify better and more accurately why students tested as they did under the direction of classroom teachers, but also how to make this measurement more equitable and fair to teachers who have had test scores fluctuate from student to subject matter and over the timespan of one year to the next, as was the case with Sherri Lederman mentioned in Chapter One. The questions were designed to investigate how a teacher's critical thinking skills were measurably connected to student test scores from three specific approaches:

- Are there measurable differences in teacher critical thinking skills that have traditionally been evaluated by district administrators?
- Is the rubric that is used and interpreted by administrators accurate, and does it give back a reliable critical thinking score measurement?
- Are the data that justify a rating of "Highly Effective" and "Effective" reliable across two different scales created by the same Pearson Corporation?

The questions were answered in ways that demonstrated a need for an interdisciplinary approach that needs to be adjusted per district demographic. The research found that a teacher's critical thinking skills are not being accurately measured because qualitative data do not provide a comprehensive explanation of where the teacher and/or student may need professional development and training as well in the subject matter. The study also found that teacher self-reporting from a psychometric perspective provides valuable data that school districts have not yet considered as adding to understanding variances in test scores. Teacher input is valuable because teachers are the ones most intimately interacting with students on a daily basis and having specialized knowledge of them. A qualitative test also does not factor in variables such as race, ethnicity, social status, and socioeconomics as part of the student and the student body profile, which can, in fact, have a significant impact on instruction and learning retention.

Evaluation systems or rating systems should come in the form of statistically sound and consistently verified and accurate measurements. Clear expectations within a multiple rating system that provides feedback in a timely manner is another important function of evaluating performance. The benefit of the Watson-Glaser II Critical Thinking Appraisal is that, besides all the information that has been mentioned, this tool is already widely used to ascertain specific work-related characteristics that will increase organization and productivity in educational institutions.

Limitations to the Study

This section discusses that limitations that exist in the apparent boundaries with which this research must contend. Such variables relate to the components within the interdisciplinary section, such as the ability of other districts strained by socioeconomic variables to offer a better model of evaluation methods.

Issues with aligning more accurate critical thinking evaluation scores may not transfer over to schools that cannot budget for professional development courses. The school district that was tested is well funded, with an average household income that is substantially higher than the state and national income. The school district illustrated a lack of cultural diversity after accessing the Report Card under the NYSED Data website retrieved through the district's homepage, which gives public access to information on population demographics (see Figure 5.1).

Because the district is not diversified, this may potentially cause a limitation to the study. It is important to note that there may be constraints on generalizability besides what was found internally from the statistical data. Applications of the Watson-Glaser II Critical Thinking Appraisal may be additionally hindered along with establishing internal and external validity.

In 2012, the average household income of the district that participated in the research was $131,212, compared to a state income average of $86,097 and a national average of $77,190 (CLRESEARCH, 2018). As Posey (2016) stated, "The U.S. Census Bureau reported in September 2016 that real median household income was $55,775 in 2015" (p. 1). With a

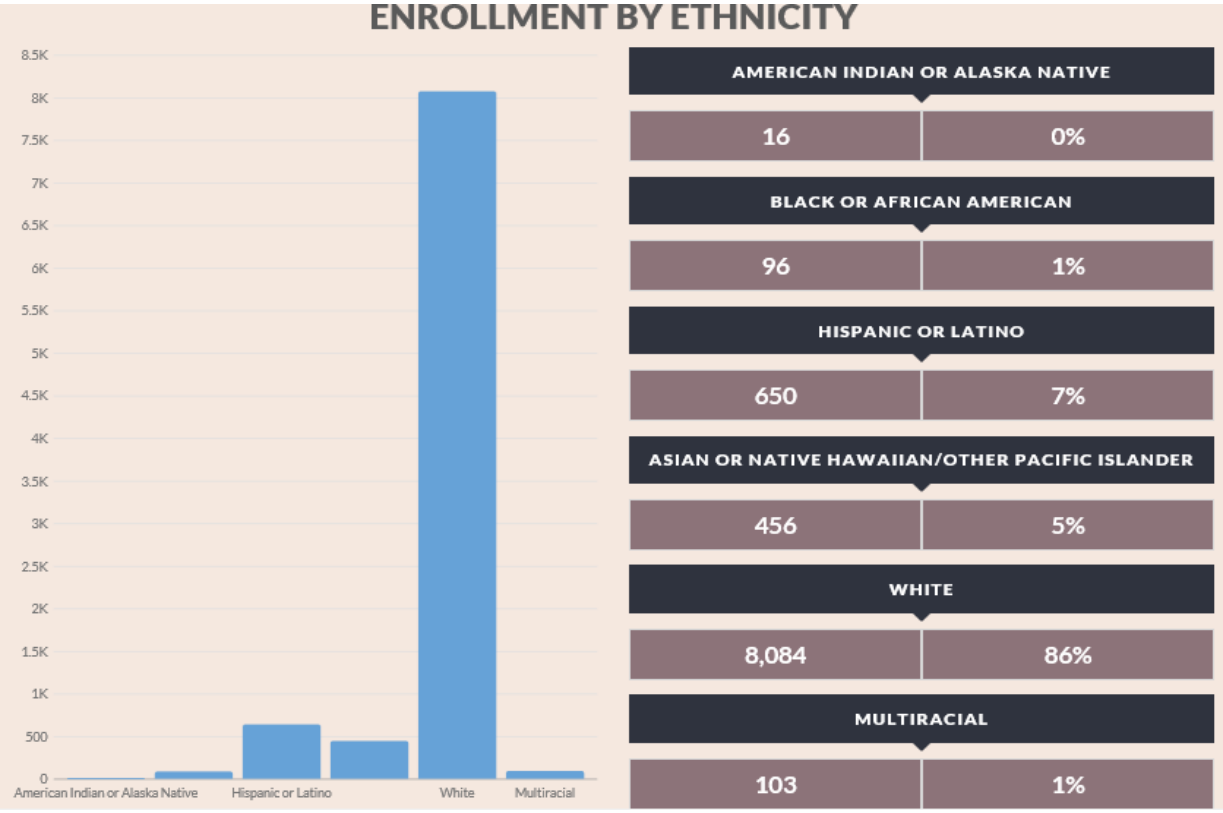


Figure 5.1. Report Card NYSED district homepage enrollment by ethnicity model

significantly large family income and school taxes, there are more resources available for ELA (English Language Arts) and ESL (English as a Second Language) programs, along with offering trained school personnel to assist student needs by involving them in special education programs and properly designing a child's IEP (Individualized Education Program). Unfortunately, other schools in low socioeconomic districts may not have sufficient funding for such programs. Further application of this study can potentially be hindered in states with lower school funding that may lack the fiscal resources to purchase self-reporting critical thinking tests. School districts with considerably lower funding will no doubt offer lower wages. Finally, the impact on the state average per pupil spending, in a report by the Census Bureau on June 2, 2015, indicated that "States and state-equivalents spending the most per pupil in 2013 were New York ($19,818), Alaska ($18,175), the District of Columbia ($17,953), New Jersey ($17,572) and Connecticut ($16,631). States spending the least per pupil included Utah ($6,555), Idaho ($6,791), Arizona ($7,208), Oklahoma ($7,672) and Mississippi ($8,130)" (para. 5).

Another concern that can emerge in this study is inter-rater reliability. Since APPR scores are established by the criteria of potentially different administrators, it will be necessary to further align how these scores are ascertained. Cohen's Kappa may be used, assuming the conditions will allow for the statistical use of this tool.

Recommendation for Future Research

Besides offering the use of such assessment tools as Watson-Glaser Critical Thinking Appraisal to substitute the APPR criteria for determining the level of "Critical Thinking" skills, other notable quantitative self-reporting scales should be investigated for use in school districts. These kinds of data will also allow evaluators with intimate knowledge of their districts to design an assessment based on the demographical makeup of the student body. They will also be able to

manage adherence to NYSUT rubric standards in a meaningful and knowledgeable way, thereby securing a foothold in meeting and staying within district and national standards. A more hands-on role in how assessment is designed also provides principals and administrators with additional data that can be used in applications for any necessary additional federal funding to help students.

Taking a quantitative approach would also help school districts accurately evaluate pedagogical standards during unannounced or announced APPR evaluations. By questioning the veracity of categorical values obtained by administrators' criteria against a more statistically reliable quantitative measure on the self-reporting Watson-Glaser Critical Thinking Appraisal, school districts can architect and revamp their system of evaluation. The value of using a quantitative tool like the Watson-Glaser Critical Thinking Appraisal is the ability to fine-tune the assessment based on the makeup of the student body, teacher experience and tenure of certification, teaching style, environmental variables, and access to teacher resources. This would include rethinking a re-articulation of the Danielson framework to incorporate these categories. This will also further determine the reliability of each school district's method of evaluating teachers and offer better statistical measures than just relying on administrator observations. Teacher perceptions pertaining to the advantages and disadvantages of the evaluation system can be assessed and compared for the prospects of increasing better transitions for evaluation implementation. Professional development courses should be researched and evaluated to increase properly and align teachers' level of critical thinking and strategies for college readiness skills in their students. Further research is necessary to analyze if critical thinking skills are demonstrated with substantial reliabilities across multiple lessons and multiple classes for a single teacher.

As another recommendation, since there is usually more than one evaluator observing teachers, more statistical analysis is needed to ensure high confidence levels toward the internal validity of inter-reliability. Engaging teachers during conference days to devote their time toward staff development in the area of improving critical thinking skills would be a necessary link for student graduation rates as well as cultivating college readiness skills and preparation. At the start of a school year, districts might find value in online instructional programs that foster critical thinking, in-service courses that are specialized for each department K-12, and workshops and three-day mentoring programs that are offered multiple times a year.

Regardless of how districts go about it, a multi-measure approach to teacher evaluations and assessment cannot be dependent on one particular method of measurement. Teachers, like their students, do not come in a "one size fits all" mold. Each district, student, teacher, and administrative office must come to a consensus on the goals of the district, all the while keeping the students and their achievements at the forefront of their minds. As the world continues to grow in technological developments and economic progress, assessment models must grow and evolve in kind. The most important perspective involved with analyzing the practical use of the New York State teacher evaluation system is the consideration made toward its ability to deliver statistically reliable results. Some considerations are being made to allow flexibility in accommodating teachers' instructional practices. The considerations include utilizing a coherent statistical method of assigning a rating score by district administrators, proposing adequate training in teacher observations, and instilling legitimate guidelines that are consistent toward ensuring equitable measures across various districts and disciplines. School boards must take the responsibility for adopting a uniform evaluation system that provides teachers with an overall rating score that reports valid, accurate, and reliable measures demonstrating high-quality

statistical data. The interdisciplinary approach should be used in a collaborative way to improve the standard by which data are collected and district evaluators and other administrative personnel are trained. If the method used to rate teachers has a large degree of reliability, then results can more effectively be communicated over periods of time throughout the year.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Achinstein, B., Curry, M., & Ogawa, R. (2015). (Re) labeling social status: Promises and tensions in developing a college-going culture for Latina/o youth in an urban high school. *American Journal of Education, 121*, 311-345. Retrieved from http://www.jstor.org.ezproxy.lib.uh.edu/stable/10.1086/680436

Adams, N. G., Shea, C. M., Liston, D. D., & Deever, B. (2005). *Learning to teach: A critical approach to field experiences*. Mahwah, NJ: Lawrence Erlbaum.

Almager, I. (2016). Harnessing white teachers' social capital: A Latina principal's perspective. *National Forum of Applied Educational Research Journal, 29*(1-2), 46-53.

Appleby, D. C. (2014, May). How do college freshmen view the academic differences between high school and college? Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago.

Arnold, K., Lu, E., & Armstrong, K. (2012). Special issue: The ecology of college readiness. *ASHE Higher Education Report, 38*(5), 1-38. doi:10.1002/aehe.20005

Aushenker, M. (2017, January 13). District seeks input on grant spending. *Simi Valley Acorn*. Retrieved from https://www.simivalleyacorn.com/articles/district-seeks-input-on-grant-spending/

Aushenker, M. (2017, January 20). Grant spending plan approved. *Simi Valley Acorn*. Retrieved from https://www.simivalleyacorn.com/articles/grant-spending-plan-approved/

Asamsama Hemmy, O., Mayo, D., Stillman, J., Mathews, C., Schnorr, D. & Nelson, B. (2016). Parental expectations, gender, and ethnicity: Predictors of college readiness. *Journal of Educational Research and Innovations, 5*(1), 1-11. Retrieved from http://digscholarship.unco.edu/jeri/vol5/iss1/3

Barnes, W. B., & Slate, J. R. (2013). College-readiness is not one-size-fits-all. *Current Issues in Education, 16*(1), 1-11. Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/1070

Bennett, G. K. (2008). Bennett mechanical comprehension test. *NCS Pearson, 32.* Retrieved from http://talentlens.pearsonpsychcorp.com.au/files/BMCT_ST_ Manual%281%29.pdf

Bernhardt, P. E. (2013). The advancement of individual determination (AVID) program: Providing cultural capital and college access to low-income students. *School Community Journal, 23*(1), 2013-2022.

The Bill and Melinda Gates Foundation. (2009). Initial findings from the measures of effective teaching project. Retrieved from https://docs.gatesfoundation.org/documents/preliminary-findings-research-paper.pdf

Bobbitt, F. (1918). *The curriculum*. Chicago, IL: Houghton Mifflin.

Bobbitt, F. (1924). *How to make a curriculum*. Chicago, IL: Houghton Mifflin.

Bok, J. (2010). The capacity to aspire to higher education: "It's like making them do a play without a script." *Critical Studies in Education, 51*(2), 163-178.

Bol, L., & Berry, R. Q. (2005). Secondary mathematics teachers' perceptions of the achievement gap. *The High School Journal, 88*(4), 32-45. doi:10.1353/hsj.2005. 0007

Brisbane, Nicole. (2015, April 16 qtd. as cited in Burman). 155,000 New York kids boycott standardized tests. *USAToday*. Retrieved from

https://www.usatoday.com/story/news/nation/2015/04/16/parents-opt-out-standardized-tests/25896607/

Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource for improvement.* New York, NY: Russell Sage Foundation.

Burman, J. (2015, April 16). 155,000 New York kids boycott standardized tests. *USAToday*. Retrieved from https://www.usatoday.com/story/news/nation/2015/04/16/parents-opt-out-standardized-tests/25896607/

Cam, P. (2000). Philosophy, democracy and education: Reconstructing Dewey. In *Teaching philosophy for democracy* (pp. 158-181). Retrieved July 29, 2017. Seoul, Korea: Seoul University Press.

Close, K., & Amrein-Beardsley, A. (2018). Learning from what doesn't work in teacher evaluation. *Phi Delta Kappan, 100*(1), 15-19.

Coggsall, J., Ott, A., & Lasagna, M. (2010). Retaining teacher talent: Convergence and contradictions in teachers' perceptions of policy reform ideas. Retrieved March 6, 2019, from https://files.eric.ed.gov/fulltext/ED508143.pdf

Conley, D. T. (2007). *Redefining college readiness,* 3. Eugene, OR: Educational Policy Improvement Center.

Daniels, E. A. (2013). APPR, solution or problem? A critical examination of education law §3012-c. Retrieved July 7, 2017, from http://www.academia.edu/9338218/APPR_solution_or_problem_A_critical_examination_of_education_law_3012-c

Danielson, Charlotte (2011). NYSUT's teacher practice rubric aligned with the New York teaching standards. Retrieved from

https://www.nysut.org/~/media/files/nysut/resources/2013/april/ted/2011_sedapproved_nysut_tpr.pdf?la=en Dake, L. (2011). Deconstructing the humanist-efficiency paradigm: Responsibly challenging the current model. *International Journal of Humanities and Social Science* (Special issue), 208-214. Retrieved July 7, 2017, from http://ijhssnet.com/journals/Vol_1_No_15_Special_Issue_October_2011/25.pdf

Darling-Hammond, L. (2014, Spring). Teacher evaluation as part of a comprehensive system for teaching and learning. Retrieved March 13, 2018, from https://www.aft.org/ae/spring2014/darling-hammond

Decker, G. (2014, December 16). 91% of city teachers rated effective or higher in first round of new evaluations. *Chalkbeat education news: In context.* Retrieved from http://www.chalkbeat.org/posts/ny/2014/12/16/91-percent-of-city-teachers-rated-effective-or-higher-on-first-evaluations/

Dressel, P. L., & Mayhew, L. B. (1954). *General education: Explorations in evaluation.* Washington, DC: American Council on Education.

Duffet, A., Farkas, S., Rotherham, A. J., & Silva, E. (2008). *Waiting to be won over: Teachers speak on the profession, unions, and reform*. Washington, DC: Education Sector.

Dunston, P., & Wilkins, J. (2015). False hope: Underprepared students' pursuit of postsecondary degrees. *English Teaching: Practice and Critique, 14*(1), 44-59. doi:10.1108/ETPC-11-2014-0002

Duerr, L. L. (2008). Interdisciplinary instruction, educational horizons. Retrieved from https://eric.ed.gov/?id=EJ798522

Edmonds, R. (1979, October). Effective schools for the urban poor. *Educational Leadership*. Retrieved from http://inservice.ascd.org/effective-schools-for-the-urban-poor/

edTPA. (2016). Annual administrative report on edTPA data shows continued growth and support for the first nationally available assessment of teacher candidates [Press release]. Retrieved from http://edtpa.aacte.org/news-area/1839.html#page

Feeney, S. (2013). An open letter of concern regarding New York state's APPR legislation for the evaluation of teachers and principals. Retrieved from http://www.newyork principals.org/appr-paper

Firestone, W., & Riehl, C. (Eds.). (2005). *A new agenda for research in educational leadership.* New York, NY: Teachers College Press.

Found, P., & Hughes, L. (2016). Understanding effective problem solving. In A. Chiarini, N. Rich, & P. Found (Eds.), *Understanding the lean enterprise: Strategies, methodologies, and principles for a more responsive organization* (pp. 130-152). New York, NY: Springer.

Fritzberg, G. (2012). A brief history of education reform: The federal government's efforts to improve our schools. Retrieved from https://spu.edu/depts/uc/response/new/ 2012-spring/features/history-of-reform.asp

Fruchter, N. M., Hester, M., Mokhtar, C., & Shahn, Z. (2012). *Is demography still destiny? Neighborhood demographics and public high school students' readiness for college in New York City.* Providence, RI: Annenberg Institute for School Reform, Brown University.

Gaertner, M., & Larsen McClarty, K. (2015). Performance, perseverance, and the full picture of college readiness. *Educational Measurement: Issues and Practice, 34*(2), 20-33.

Glazerman, S., Loeb, S., Goldhaber, D., Raudenbush, D., Staiger, D., & Whitehurst, G. J. (2010). *Evaluating teachers: The important role of value-added.* Washington, DC: The Brookings Brown Center.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved March 3, 2009, from http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

Gravetter, F., & Wallnau, L. (2014). *Essentials of statistics for the behavioral sciences*. Belmont, CA: Wadsworth Cengage Learning.

Harris, E. (2016, May 11). Court vacates Long Island teacher's evaluation tied to test scores. *The New York Times,* Sec. A, p. 21.

Harvey, D., Slated, J., Moore, G., Barnes, W., & Martinez-Garcia, C. (2013). College readiness gaps: A review of the literature. *Journal of Education Research, 7*(3), 181-204.

Hernandez, J. C. (2017, July 31). Study finds Chinese students excel in critical thinking until college. *New York Times.* Retrieved from http://www.nytimes.com/2016/07/31/world/asia/china-college-education-quality.html?_r=0

Hernandez, K. (2017, November 17). How Redlands unified grads can be guaranteed University of Redlands college admission. Retrieved from https://www.redlandsdailyfacts.com/2016/11/17/how-redlands-unified-grads-can-be-guaranteed-university-of-redlands-admission/

Hiltensmith, R., & Draut, T. (2014, March). The great cost shift continues: State higher education funding after the recession, 1-13. Retrieved from demos.org.

Hood, J. (1993). The failure of American public education. Retrieved from https://fee.org/articles/the-failure-of-american-public-education/

Hsiao, A. (2011, November 8). Race to more ineffective ed spending. *National Review*. Retrieved from http://www.nationalreview.com/corner/282565/race-more-ineffective-ed-spending-annie-hsiao

Jackson, J., & Kurlaender, M. (2014). College readiness and college completion at broad access four-year institutions. *American Behavioral Scientist, 58*(8), 947-971. doi:10.1177/0002764213515229

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136. doi:10.1086/522974

Jones, C. (2009). Interdisciplinary approach—Advantages, disadvantages, and the future benefits of interdisciplinary studies. *ESSAI,* 7, Article 26. Retrieved from https://dc.cod.edu/cgi/viewcontent.cgi?article=1121&context=essai

Jones, V. (2016, March). Social and cultural capital: The heart of STEM success. *Children's Technology and Engineering,* 16-19.

Karp, S. (2013, Winter). The problems with the Common Core. *Rethinking Schools, 28*(2). Retrieved from http://www.rethinkingschools.org/ProdDetails.asp?ID= RTSVOL28N2

Kennedy, M. (2010). The uncertain relationship between teacher assessment and teacher quality. In M. M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook.* San Francisco, CA: Jossey-Bass.

Kennedy, M., Fisher, M. B., & Ennis, R. H. (1991). Critical thinking: Literature review and needed research. In L. Idol & B. F. Jones (Eds.), *Educational values and cognitive instruction: Implications for reform* (pp. 11-40). Hillsdale, NJ: Lawrence Erlbaum.

Kimball, S., White, B., & Milanowski, A. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54.

Kliebard, H. (2004). *The struggle for the American curriculum: 1893-1958*. New York, NY: Routledge Falmer.

Koch, B., Slate, J., & Moore, G. (2012, Fall). Perceptions of students in developmental classes. *Community College Enterprise*, 62-82.

Krumrei-Mancuso, E., Newton, F., Kim, E., & Wilcox, D. (2013). Psychosocial factors predicting first-year college student success. *Journal of College Student Development, 54*(3), 247-266. doi:10.1353/csd.2013.0034

Lai, E. (2011). *Critical thinking: A literature review*. Pearson. Retrieved from http://images.pearsonassessments.com/images/tmrs/CriticalThinkingReviewFINAL.pdf

Lata, L. (2013). Cultural capital and higher education in Bangladesh. *International Review of Modern Sociology, 39*(2), 225-238.

Leal, F. (2015, July). Survey: Most high school students feel unprepared for college, careers. *EdSourc*. Retrieved from https://edsource.org/2015/survey-most-high-school-students-feel-unprepared-for-college-careers/83752

Lillian Moller Gilbreth—The first lady of management. (2017). Retrieved July 24, 2017, from https://www.kaplanuniversity.edu/news-resources/first-lady-of-management/ Kaplan Higher Education, LLC.

Maharaj, S. (2014). Administrators' view on teacher evaluation. *Canadian Journal of Educational Administration and Policy, 152,* 1-58. Retrieved July 25, 2018, from https://eric.ed.gov/?id=EJ1021929

McCaffrey, D., Lockwood, J. R., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models of teacher accountability.* Santa Monica, CA.; Rand Corporation.

Nagaoka, J., Farrington, C., Roderick, M., Allensworth, E., Keyes, T., Johnson, D., & Beechum, N. (2013, Fall). Readiness for college: The role of non-cognitive factors and context. *Vue,* 45-52.

National Center for Education Statistics. (2016). *The condition of education 2016: Undergraduate retention and graduation rates.* (NCES 2016-144). Washington, DC: Author.

New York State Department of Education (NYSED). (2011). Guidance on New York State's annual professional performance review law and regulations. Retrieved April 17, 2018, from https://eric.ed.gov/?id=ED524871

New York State teacher's rubric of teaching standards (2012). *New York State United Teachers*. Retrieved from https://usny.nysed.gov/rttt/teachers-leaders/practice rubrics/Docs/nysut-rubric-2012.pdf

Nie, N., Bent, C., & Hull, H. (1970). *SPSS: Statistical package for the social sciences.* New York, NY: McGraw-Hill.

Odden, A., Borman, G. & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4-32. doi:10.1207/s15327930pje7904_2

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Pellmar, T. C., & Eisenberg, L. (Eds.). (2000). The potential of interdisciplinary research to solve problems in the brain, behavioral, and clinical sciences. *Institute of Medicine (US) Committee on Building Bridges in the Brain.* Washington, DC: National Academies Press.

Person, A., Baumgartner, S., Hallgren, K., & Santos, B. (2014, September). Measurement and segmentation of college students' noncognitive attributes: A targeted review. *Mathematica Policy Research,* 1-29.

Posey, K. G. (2016, September 16). Household income: 2015. Retrieved May 20, 2017, from https://www.census.gov/library/publications/2016/acs/acsbr15-02.html. Report Number: acsbr/15-02

Ramaswamy, S.V. (2015, April 16). 155,000 *New York kids boycott standardized tests. USA Today*. USATODAY. https://www.usatoday.com/story/news/nation/2015/04/16/parents-

RAND Corporation. (2012). Multiple choices: Options for measuring teacher effectiveness. Retrieved April 14, 2017, from RAND Corporation website: http://www.rand.org/education/projects/measuring-teacher-effectiveness/multiple-choices.html

Repko, A. F., Szostak, R., & Buchberger, M. P. (2014). *Introduction to interdisciplinary studies*. Thousand Oaks, CA: Sage.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175-214.

Sagnak, M. (2010). The relationship between transformational school leadership and ethical climate. *Educational Sciences: Theory and Practice, 10*(2), 1135-1152.

Schaffhauser, D. (2015, July). Survey: Most profs find HS grads unready for college or work. *Campus Technology*. Retrieved from https://campustechnology.com/articles/2015/07/27/survey-most-profs-find-hs-grads-unready-for-college-or-work.aspx

Schwartz, N. (2013). *Survey captures teachers' perception of evaluation system*. Retrieved from http://tnclassroomchronicles.org/survey-captures-teacher-perceptions-of-evaluation-system/

Smylie, M. A. (2014). Teacher evaluation and the problem of professional development. In B. Superfine (Ed.), *Research on urban education policy initiative*. Chicago, IL: University of Illinois at Chicago.

Springer, S., Wilson, T., & Dole, J. (2014). Ready or not: Recognizing and preparing college-ready students. *Journal of Adolescent and Adult Literacy, 58*(4), 299-307.

Staples, H. (2005). The integration of biomimicry as a solution-oriented approach to the environmental science curriculum for high school students. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/c2/3d.pdf

Strauss, V. (2015, August 9). Master teacher suing New York State over 'ineffective' rating is going to court. *New York Post*. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2015/08/09/master-teacher-suing-new-york-state-over-ineffective-rating-is-going-to-court/?utm_term=.9103dd125d6f

Stringer (Comptroller), S. M. (2016). *Diploma disparities: High school graduation rates in New York City*. Retrieved from Office of the New York City Comptroller website: https://comptroller.nyc.gov/wp-content/uploads/documents/Graduation_Rate_ Brief.pdf

Taylor, F. W. (2014). *The principles of scientific management.* Eastford, CT: Martino Fine Books, 2014.

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education.* Washington, DC: Education Sector.

Tyrell, J. (2016). Tens of thousands of LI students opt out of common core exams. *Newsday*. Retrieved from https://www.newsday.com/long-island/education/tens-of-thousands-of-li-students-opt-out-of-common-core-exams-1.11653222

Urban Opt Out Statistics in NY State (2017). *New York City opt out.* Retrieved from https://www.optoutnyc.com/urban-opt-out-stats

U. S. Department of Education (USDOE). (2009). *The facts about No Child Left Behind*. Retrieved from http://www2.ed.gov/nclb/overview/intro/parents/parentfacts.html

U. S. Department of Education (USDOE). (2014). *Setting the pace: Expanding opportunity for America's students under Race to the Top*. Retrieved from http://www.white house.gov/sites/default/files/docs/settingthepacerttreport_3-2414_b.pdf

VanSickle, R. L. (2012). Research implications of a theoretical analysis of John Dewey's *How We Think. Theory and Research in Social Education, 13*(3), 1-20. doi:10.1080/00933104.1985.10505503

Wadsworth, B. J. (2004). *Piaget's theory of cognitive and affective development* (5th ed.). Upper Saddle River, NJ: Pearson. doi:0205406033, 9780205406036

Webb, M., & Thomas, R. (2015). Teachers' perceptions of educators and students' role in closing the achievement gap. *National Forum of Teacher Education Journal, 25*(3), 1-8.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* Retrieved February 1, 2010 from the New Teacher Project website: http://www.tntp.org/

Will, M. (2018). An expensive experiment: Gates teacher effectiveness program shows no gains for students. *Education Week, 37*(37), p. 9.